ROWAN GILMORE • LES BESSER

PRACTICAL RF CIRCUIT DESIGN FOR MODERN WIRELESS SYSTEMS

VOLUME II

ACTIVE CIRCUITS AND SYSTEMS

VISIT…

Practical RF Circuit Design for Modern Wireless Systems

Volume II

Active Circuits and Systems

For a listing of recent titles in the *Artech House Microwave Library*,
turn to the back of this book.

Practical RF Circuit Design for Modern Wireless Systems

Volume II

Active Circuits and Systems

Rowan Gilmore
Les Besser

Artech House
Boston • London
www.artechhouse.com

Library of Congress Cataloging-in-Publication Data
Gilmore, Rowan.
Practical RF circuit design for modern wireless systems/Rowan Gilmore, Les Besser.
v. cm.— (Artech House microwave library)
Includes bibliographical references and index.
Contents: v. 2. Active circuits and systems
ISBN 1-58053-522-4 (v. 2: alk. paper)
1. Radio circuits—Design and construction. 2. Microwave circuits—Design and construction. 3. Wireless communication systems—Equipment and supplies.
I. Besser, Les. II. Title. III. Series.

TK6560.G45 2003
621.384'12—dc21 2003048107

British Library Cataloguing in Publication Data
Gilmore, Rowan
Practical RF circuit design for modern wireless systems
Vol. 2: Active circuits and systems.—(Artech House microwave library)
1. Radio circuits—Design 2. Wireless communication systems
I. Title II. Besser, Les
621.3'8412
ISBN 1-58053-522-4

Cover design by Yekaterina Ratner. Text design by Darrell Judd.

©2003 ARTECH HOUSE, INC.
685 Canton Street
Norwood, MA 02062

All rights reserved. Printed and bound in the United States of America. No part of this book may be reproduced or utilized in any form or by any means, electronic or mechanical, including photocopying, recording, or by any information storage and retrieval system, without permission in writing from the publisher.

All terms mentioned in this book that are known to be trademarks or service marks have been appropriately capitalized. Artech House cannot attest to the accuracy of this information. Use of a term in this book should not be regarded as affecting the validity of any trademark or service mark.

International Standard Book Number: 1-58053-522-4
Library of Congress Catalog Card Number: 2003048107

10 9 8 7 6 5 4 3 2 1

To our wives, Nicole and Susan

Contents

Preface xiii

Acknowledgments xvii

1 Linear RF amplifier design— general considerations 1

1.1 Introduction 1
1.2 Power gain definitions 3
1.3 Neutralization 7
1.4 Unilateral transducer gain 8
1.4.1 Unilateral figure of merit 10
1.4.2 Illustrative example: unilateral gain calculations 12
1.4.3 Amplifier design with single matching networks 13
1.4.4 Unilateral constant gain circles 15
1.4.5 Illustrative example: single-sided amplifier design 15
1.5 RF circuit stability considerations 19
1.5.1 What may cause RF oscillation 22
1.5.2 Stability analysis with arbitrary source and load terminations 25
1.5.3 Two-port stability considerations 30
1.5.4 Stability circles 35
1.5.5 Graphical forms of unconditional stability 40
1.5.6 Graphical forms of potential instability 41
1.5.7 Caution about multistage systems 42
1.6 Stabilizing an active two-port 46
1.6.1 Finding the minimum-loss resistor at the input of the device 47
1.6.2 Broadband stability considerations 49
1.7 Stabilization of a bipolar transistor 50
1.7.1 Examining the effect of lossless feedback 50
1.7.2 Device stabilization 51
1.8 The dc bias techniques 59
1.8.1 Passive dc bias networks 60
1.8.2 Active dc bias circuits 63
1.8.3 Feeding dc bias into the RF circuit 64
1.8.4 The dc bias circuit simulation 65
1.8.5 Filtering of dc bias networks 69
1.9 Statistical and worst-case analyses 69

1.10 Circuit layout considerations 71
1.11 Summary 74
1.12 Problems 74
References 75
Selected bibliography 76

2 Linear and low-noise RF amplifiers 77

2.1 Introduction 77
2.2 Bilateral RF amplifier design for maximum small-signal gain 78
2.2.1 Illustrative exercise: amplifier design for maximum gain, G_{MAX} 82
2.3 Multistage amplifiers 88
2.3.1 Cascading impedance-matched stages 88
2.3.2 Cascading amplifiers by direct impedance matching 89
2.3.3 Output power and impedance match considerations of cascaded amplifiers 92
2.4 Operating gain design for maximum linear output power 94
2.4.1 Operating gain design outline 95
2.4.2 G_p versus P_{OUT} trade-offs 97
2.4.3 Stability considerations 97
2.4.4 Illustrative example: operating gain design for maximum linear power output 98
2.4.5 Output match considerations 101
2.5 Noise in RF circuits 102
2.5.1 Review of noise sources in RF systems 102
2.5.2 Two-port noise parameter definitions 106
2.6 Available gain design technique 107
2.6.1 Available gain design outline 108
2.6.2 Low-noise amplifier design considerations 110
2.6.3 Illustrative example: design of a single-ended 1.9-GHz LNA 111
2.6.4 Balanced amplifiers 114
2.6.5 Illustrative example: design of a balanced LNA for the 1.7- to 2.3-GHz frequency range 116
2.7 Comparison of the various amplifier designs and Smith chart–based graphical design aids 121
2.8 Broadband amplifiers 123
2.8.1 Reactive match/mismatch approach 124
2.8.2 Dissipative mismatch at input and/or output ports 125
2.8.3 Amplifier-equalizer combinations 129
2.8.4 Feedback amplifiers 129
2.8.5 Distributed amplifiers 141
2.9 Summary 142
2.10 Problems 143
References 144
Selected bibliography 145

3 Active RF devices and their modeling 147

3.1 The diode model 148
3.2 Two-port device models 150
3.2.1 The output terminals of a two-port RF device 150
3.2.2 The bipolar transistor 153
3.2.3 The heterojunction bipolar transistor 173
3.2.4 The GaAs MESFET 177
3.2.5 The high-electron mobility transistor 184
3.2.6 Silicon LDMOS and CMOS technologies 187
3.3 Problems 190
References 190

4 Nonlinear circuit simulation techniques 193

4.1 Classification of nonlinear circuit simulators 193
4.1.1 Analytical methods 194
4.1.2 Time-domain methods 194
4.1.3 Hybrid time- and frequency-domain techniques—harmonic balance 197
4.1.4 Frequency-domain techniques 200
4.2 The harmonic balance method 202
4.3 Harmonic balance analysis of oscillators 207
4.3.1 Oscillator analysis using probes 208
4.3.2 Oscillator analysis using reflection coefficients of the device and resonant load 209
4.3.3 Oscillator analysis using a directional coupler to measure open-loop gain 214
References 215

5 High-power RF transistor amplifier design 217

5.1 Nonlinear concepts 217
5.1.1 Some nonlinear phenomena 220
5.2 Quasi-linear power amplifier design 223
5.2.1 The amplifier load line 224
5.2.2 Load pull methods 232
5.3 Categories of amplifiers 243
5.3.1 Class-A amplifier 243
5.3.2 Class-B amplifier 248
5.3.3 Class-F amplifier 257
5.3.4 Comparison of class-A, class-B, class-F, and other operational modes 265
5.3.5 Switching-mode amplifiers 271
5.3.6 Cascaded power amplifier design 278

5.4 Power amplifier design example 280
5.4.1 Transistor selection 281
5.4.2 Transistor characterization 282
5.4.3 Matching the input and output of the device 286
5.4.4 Harmonic tuning example 296
5.5 Bias considerations 298
5.5.1 Bias changes at the input 298
5.5.2 Bias changes at the output 302
5.5.3 Bias considerations with power devices 304
5.6 Distortion reduction 307
5.6.1 The importance of amplifier linearity 309
5.6.2 Operating the amplifier backed off 311
5.6.3 Predistortion 312
5.6.4 Feedforward cancellation 317
5.6.5 Device modification 319
5.6.6 System-level reduction of distortion 325
5.7 Problems 328
References 334

6 Oscillators 337

6.1 Principles of oscillator design 338
6.1.1 Two-port oscillator design approach 338
6.1.2 One-port oscillator design approach 349
6.1.3 Transistor oscillator configurations 373
6.1.4 Characterizing oscillator phase noise 390
6.2 Oscillator design examples 404
6.2.1 45.455-MHz Colpitts crystal oscillator design 404
6.2.2 Design of a 3.7- to 4.2-GHz voltage-controlled oscillator 410
6.3 Problems 429
References 431

7 Mixers and frequency multipliers 433

7.1 Mixer overview and their applications in systems 433
7.2 Diode mixers and their topologies 442
7.2.1 Single-ended mixer 443
7.2.2 Single-balanced mixer 445
7.2.3 Double-balanced mixer 451
7.2.4 The image problem in mixers 455
7.2.5 Harmonic components in mixers 460
7.3 Transistor mixer design 464
7.3.1 Active transistor mixers 464
7.3.2 Resistive FET mixers 488
7.3.3 Dual-gate FET mixers 494
7.3.4 Comparison of mixers 500

7.4 Frequency multipliers—an overview 501
7.4.1 Frequency doublers 502
7.4.2 Arbitrary frequency multiplication 505
7.5 Problems 506
References 507

8 Circuits in systems—radio system applications 509
8.1 Mobile telephony systems 509
8.1.1 Second generation mobile systems 510
8.1.2 Third generation mobile systems 512
8.2 Software-defined radio 515
8.2.1 RF digital processing 515
8.2.2 Digital processing of a wideband IF 517
8.2.3 Digital processing at baseband (direct conversion) 518
8.2.4 Transceiver issues associated with software-defined radio 520
8.3 A 1.9-GHz radio chip set: design overview 522
8.3.1 The air interface specification for PHS 522
8.3.2 Component specification 523
8.3.3 Component design 525
8.4 Integrated system chips: an overview 531
8.4.1 RF receiver front ends 532
8.4.2 RF upconverters and transmitter driver amplifiers 536
8.4.3 Transceiver and complete radio solutions 538
8.4.4 Power amplifier modules 543
8.5 Conclusion 544
References 545

Appendix 547
Summary of Basic Formulas – 1 547
Summary of Basic Formulas – 2 549

About the Authors 551

Index 553

Preface

This text is intended to be a populist book.

With so many complex equation–filled engineering books lining the shelves of our bookstores, perhaps you are wondering whether the science of microwaves and RF is ready for a text that can be understood by those who do not speak Latin or wear black robes. We believe so. The goal of a populist book is to appeal as much to the academic at a highbrow university as to the practitioner working in today's frantic production environment. We hope you will find this text as relevant to your work of teaching others as to improving your own skills.

This book is written for practicing engineers and for those who would like to become one. And these days, who can afford not to keep learning? Whether you are a student at your final year of college, an engineer in industry who has just been assigned your first RF design project, or a seasoned veteran of the magic of microwave design, we hope that you will all find something useful in these pages. Even if you are a microwave or RF industry guru with most of the answers already, our experience in writing this has been that there is still a thing or two out there that needs explaining. If you cannot find anything that seems inexplicable, then at least you will have the satisfaction of reassuring yourself that you have indeed been right all these (long!) years.

We do not suggest you throw away your other excellent text books that explain semiconductor transport equations, Green's functions, or the complex mathematics of filter design; just that this effort might make those paperweights all the more relevant. Do not misunderstand us—we do not imply that anyone can become a high-grade RF circuit and system designer without using any complex algebra. We feel strongly, however, that you *do not need as much* of it as some of the courses you have taken before may have included.

This book and Volume I are the culmination of more than 40 joint years of teaching these topics to *thousands of practicing electrical engineers* from around the world. Little by little, we have extended the scope of our courses and learned the simplest ways to convey basic ideas to our audience. We have often been surprised and have found for the most part that our audience is generally not interested in obtaining guru status or academic knowledge, but interested rather in gaining an understanding of

microwave and RF circuits, in gaining intuitive insight, and in applying that to their work. We hope we have captured that spirit herein.

This book is not written for the expert. If anything, we have omitted specialist material (it is long enough as it is!). We often begin our courses by telling our students that if they have spent the past year characterizing the intermodulation properties of a device to design a predistorter circuit, they are probably already one of just a handful of experts in the world in that area—and they can probably teach us something. Although we hope this book will convey the background and insight to set you on the road to becoming an expert, it will not take you down the narrow and winding lanes that make you one. We have focused on discrete circuits and discrete circuit design rather than IC design, believing that only when discrete design is mastered can those techniques be applied to integrated circuits. In consciously stopping short of IC design, we have not considered many worthy topics, such as RC or AGC oscillators or complex biasing techniques. Nor have we considered integrated systems such as phase-locked loops. All these topics are worthily covered elsewhere in expert texts of their own, and rightly so. Perhaps a third volume of this series will one day attempt to simplify those topics as well, should our wives ever let us back near our computers again!

These two volumes can be used as a final-year text in applied RF engineering towards a bachelor's degree in electrical engineering, or as part of a master's degree coursework material, or as a reference by the engineer who has already reached that level. In the university context, it is suited for a two-semester course. We assume that the student already has an understanding of basic topics such as phasors, electromagnetics, Fourier transforms, circuit analysis, and semiconductors. To be on the safe side, Chapters 1 and 2 of Volume I summarize most of the fundamentals needed as a foundation. From that background, we recommend that the text be taught in the order in which we have presented the chapters. Our experience is that after some initial preliminaries, the systems material starting in Chapter 3 of Volume I can motivate the rest. It contains simple applications of radio technology so that the student can feel worthwhile accomplishment early on and will see good reason to pursue his or her subsequent detailed work. We attempt to close the circle at the end of this volume by returning to the radio systems aspects started at the beginning, but now armed with a more detailed understanding of the technology. Prototyping a radio system with some of the integrated circuits in the final chapter would be a worthwhile student project that could proceed throughout the year, building on the self-discovery process in parallel with the formal learning. In the middle, we cover all the important techniques of RF, such as impedance matching, device characterization and modeling, amplifiers, oscillators, and mixers. Knowing how to build high-speed blocks for gain, loss, frequency conversion, and oscillation enables the student to go on to build almost any RF component.

This text differs from others in that we focus on the systems aspects of a design. To use a metaphor, we look at the forest rather than the trees, although we have included plenty of different greenery and spend ample time examining the leaves and branches as well. We assume the student comes with a basic knowledge of agriculture, understanding the soil, rain, sun, and so on. We have not focused on any one particular topic, although because the property of amplification is so fundamental, it is covered in rather more detail. We use amplification to learn about devices, simulation, distributed elements, characterization, impedance matching, stability, gain and power, and nonlinear behavior. We have also been generous in the use of the simulator to illustrate each tree with many examples, and we encourage the student to develop his own. Our goal is that by the end of the text, the student will be able to plan and seed his or her own forest with enough interest to make it grow.

Throughout the book, we emphasize *computer-aided design* (CAD) techniques and encourage you to use them as much as possible in your daily work. At the same time, we abhor the idea of *blindfold optimization* without first obtaining a reasonable initial estimate and intuitive feel for the outcome. Although today's CAD tools are powerful enough to reach a solution at times for simple problems, relying on optimization without understanding the underlying circuit or system fundamentals is a poor practice that inevitably leads to failure. Combining CAD techniques with a thorough understanding of RF fundamentals and use of traditional engineering tools is the best way to be successful.

The text contains material that is both mature and state of the art, although we have been inclined to retain mature material if it is still current and where it can provide more fundamental understanding than a result just published in a recent journal. We believe that great textbooks are written to last for years. They should teach fundamental principles that can be applied to each recent technological advance as it comes along and not become obsolete in the process. Our courses have attempted to do just that, and in encapsulating the core of what we have taught, we hope we will achieve that here as well.

Acknowledgments

Like many others, this book has grown out of course notes. In particular, the courses Applied RF Techniques I, Applied RF Techniques II, and Applied Wireless and Microwave Techniques we have taught at Besser Associates and CEI Europe have been especially fruitful in this regard. This book, however, does differ from similar books in that we have taught these courses to the rather more demanding audience of graduate, practicing engineers, which we hope will make it quite relevant. Accompanying us in our teaching of these courses have been other notables in the industry, who have helped shape these notes, and therefore this book. In particular, they include Ed Niehenke and Alan Podell, who have helped to develop, structure, and formulate these presentations over the years. Their contributions to this book are sometimes implicit, but nevertheless manifold, especially in this volume. Other Besser instructors—Rick Fornes, Steve Hamilton, and Lynne Olsen—have also provided many helpful suggestions and improvements. The comments of Giora Goldberg, Bob Morrow, and Irving Kalet in their particular areas of expertise have also enriched the text.

We are also indebted to the numerous CAD vendors whose products we have liberally used both in this text and our coursework. In particular, the efforts of the highly professional and responsive staff at Applied Wave Research have enabled much of this work. Dane Collins, vice president of engineering at AWR, has frequently gone out of his way to assist us, as has their support group. Steve Maas, CTO at AWR, has also assisted us with technical issues.

We are also thankful to Bill Bridges, Joe Civello, Danielle Flint, Ian Piper, Sid Seward, and Al Ward at Agilent Technologies for providing circuit and device information, as well as software, so that we could keep the focus of this book practical. The contribution of Bob Stengel and Bill Eisenstadt of the University of Florida on mixed-mode *S*-parameter techniques was particularly helpful since there is little published material on the subject.

Bill Gonzalez of the University of Miami has also made many helpful suggestions and corrections to the books.

Others who assisted with design and CAD information include Rich Carlson at Motorola, Aki Nakatani at Ansoft, Mike Meehan at Agilent Technologies, Shawn Carpenter, Jim Rautio, and Volker Muehlhaus at

Sonnet, and former colleague Peter Sturzu. We acknowledge and appreciate their help.

Thanks should also go to the many staff at the semiconductor manufacturers who helped us with device information and drawings. Their companies include Agilent, Analog Devices, Anadigics, California Eastern, IBM, Infineon, M/A-COM, Maxim, Motorola, Peregrine, and Philips. We also received support from the staff of ATC, Coilcraft, Maury Microwave, Modelithics, and Murata, who helped us with their datasheets and device data—sometimes hot off the press.

Colleagues at Besser Associates, Jeff Lange and Annie Wong, have also assisted. Rex Frobenius worked long hours to make corrections and provided ideas as well as support with many of the illustrations. Les' daughter Daphne also lent her graphics expertise to generate many of our illustrations. Thanks should also go to Les' son Kent and daughter Nanci for their proofreading and helpful suggestions.

The staff at Artech House who produced this book have also worked tirelessly. Mark Walsh, Barbara Lovenvirth, Rebecca Allendorf, Judi Stone, and Darrell Judd have all managed the idiosyncrasies and extra effort of working with two authors on opposite sides of the globe. We thank them all that through their efforts we managed to finally make it to print. Our reviewers helped to reshape sections and clarify points when we had strayed, and they encouraged us when we were on the right path. We thank them for keeping us honest.

To our former students from various countries we express our gratitude and appreciation for providing criticism and suggestions about the content of the courses we have presented. We appreciated the feedback that enabled us to fine-tune our delivery, tweak our examples, and stay focused on a practical theme.

Finally and most importantly we thank our families who missed us for weeks on end as we stayed glued to our computers. There must be a fundamental law defining how much time it takes to write a two-volume book, but instead of listening to those who have done it before, we set out to prove that we could do it in 6 months—more or less, we thought. As it turned out, it was much, much more! We eventually stopped keeping track of the 18-hour workdays of thinking, writing, talking, simulating, and drawing that required withdrawal and silence. We are indebted to our families for putting up with all that. Without their understanding that *what a man's got to do, he's got to do*, this would never have been accomplished.

CHAPTER 1

Linear RF amplifier design—general considerations

Linear *radio frequency* (RF) amplifiers provide the foundation for active circuit design. From the fundamental concepts of amplifier design, we can develop an increasingly detailed understanding of active circuits. We will move from simplified design methods to powerful mathematical techniques requiring the assistance of *computer-aided design* (CAD) tools. Linear techniques will lead us to nonlinear principles, which, in turn, will enable an understanding of RF oscillators and power amplifiers.

1.1 Introduction

In Volume I we focus on passive circuits and components. Passive two-port circuits do not require dc bias and cannot increase the level of any applied input signal. When characterized by impedance parameters, the real parts of the impedances are always found in the right-hand plane (i.e., resistances have positive values). If *S*-parameters are used for the characterization, their magnitudes are always equal to or less than unity.

Active circuits and components, on the other hand, require dc bias to function properly, can provide gain, and may even generate desired and undesired signals at various frequencies. They may display reflection coefficients with magnitudes larger than unity, which represent negative resistance. Their characterization and modeling can be quite difficult, requiring more care, patience, and experience than what was needed for passive components.

In Chapters 1 and 2 of this volume, we will simply use measured two-port *S*-parameters for the active devices, rather than component models. This technique works well for amplifier design in the small-signal, linear, steady-state mode. Then we introduce active component models and use them in the rest of the book for amplifiers and other types of active circuits, such as oscillators, mixers, and power amplifiers.

Since the mathematics of active circuit design are too complex for manual computations, we rely heavily on CAD techniques, using

commercially available circuit simulators. Still, to learn enough of the underlying theory, we include important formulas and expressions—without derivations and proofs. References are provided throughout for those interested in more detailed mathematical treatment.

Classical *S*-parameter amplifier design [1, 2], which emerged in the mid-1960s, was often based on the unilateral assumption ($|s_{12}| = 0$), where matching networks connected to the input and output ports had no direct effect on each other. Under this assumption, the gain roll-off of the active device is compensated reactively by matching and mismatching the device at various frequencies to obtain flat gain response (see Figure 1.1). The appropriate circuit transformations to provide the necessary load and source terminations were then determined graphically by Smith chart manipulations. Circuit optimization helped to change the component values until the desired performance was reached.

Unilateral design[1] and optimization offered simplicity, but the technique was neither accurate nor reliable. Bilateral techniques[2] that followed required more work but they led to exact solutions. As CAD tools became more available and accepted, the unilateral approach was quickly replaced by the newer bilateral methods.

Modern CAD techniques [3–6], combined with component modeling and sound engineering judgment, have changed linear RF amplifier design from an art to a science. When suitable *S*-parameters are available for the active device, the initial small-signal RF design may be reduced to one of three bilateral *S*-parameter procedures [7], based on the following power gain expressions:

- *Transducer power gain*, for simultaneously conjugate matched input and output ports, which leads to maximum small-signal gain;

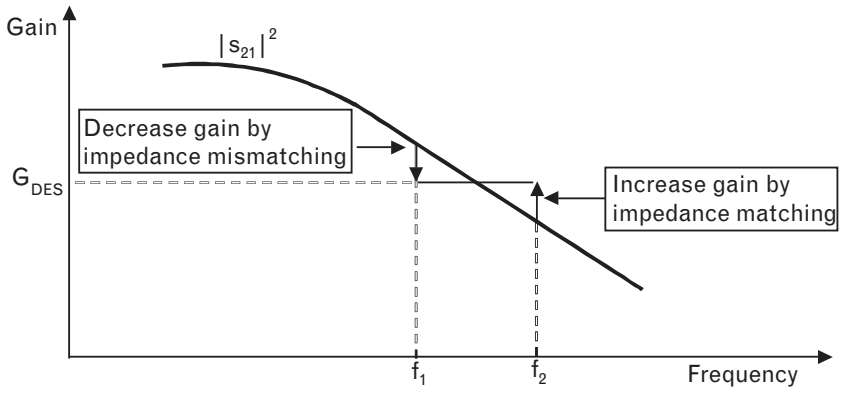


FIGURE 1.1
One possible method of amplifier design is to mismatch and match the active device at frequencies f_1 and f_2 to maintain a desired gain level G_{DES}.

1. Setting $|s_{12}|$ to zero, assuming the device has only forward transmission.
2. Including the effects of s_{12}.

- *Available power gain*, for low noise (LNA);
- *Operating power gain*, for maximum linear output power.

In this chapter we first define the necessary two-port gain expressions, and review the unilateral *S*-parameter technique and its related graphical design aids. Next, we examine RF circuit stability and device stabilization methods. Finally, we will look at dc biasing techniques and circuit layout considerations. In Chapter 2 we will set up step-by-step CAD procedures for these three amplifier categories.

1.2 Power gain definitions

We recall our definition for transducer power gain, (4.17) from Volume I, associated with the two-port network setup shown in Figure 1.2,

$$G_T = \frac{\text{Power delivered to the load}}{\text{Power available from matched source}} = \frac{P_L}{P_{AVS}} = \frac{(1-|\Gamma_S|^2)\,|s_{21}|^2\,(1-|\Gamma_L|^2)}{|(1-s_{11}\Gamma_S)(1-s_{22}\Gamma_L)-s_{12}s_{21}\Gamma_S\Gamma_L|^2} \tag{1.1}$$

where Γ_S and Γ_L are the reflection coefficients of the source and load termination, respectively. The four *S*-parameters refer to the basic Z_0 characterization[3] of the two-port. When we use the expression basic gain, or basic transducer gain, we refer to transducer gain measured with Z_0 terminations.

It is hard to find a physical interpretation to (1.1). We can develop, however, two alternative forms by introducing two new terms first, Γ_{IN} and

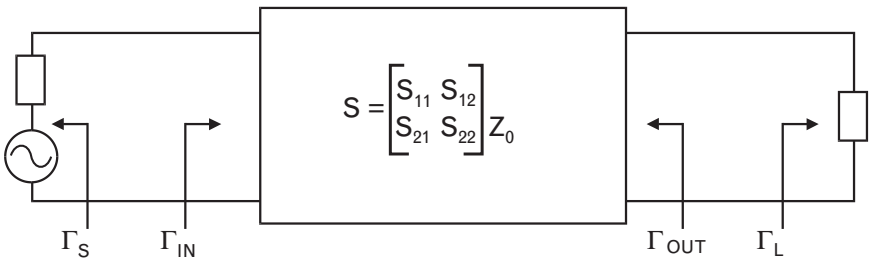


FIGURE 1.2 *Generalized block diagram of a two-port connected to arbitrary source and load terminations. The two-port is characterized by its basic Z_0-based S-parameters. In most cases the reference characteristic impedance, Z_0, is 50Ω.*

3. Measured with source and load being equal to Z_0, generally 50Ω.

Γ_{OUT}, for the input and output reflection coefficients of the two-port. Then we can rewrite (1.1) in two new forms, which have more physical meanings. Both of the new equations, (1.2) and (1.4), have three parts, representing the effects of any arbitrary source and load termination and the basic gain of the device. The first new form is

$$G_T = \frac{1-|\Gamma_S|^2}{|1-\Gamma_{IN}\Gamma_S|^2}|s_{21}|^2\frac{1-|\Gamma_L|^2}{|1-s_{22}\Gamma_L|^2} \tag{1.2}$$

where

$$\Gamma_{IN} = s_{11} + \frac{s_{12}s_{21}\Gamma_L}{1-s_{22}\Gamma_L} \tag{1.3}$$

The second new expression is

$$G_T = \frac{1-|\Gamma_S|^2}{|1-s_{11}\Gamma_S|^2}|s_{21}|^2\frac{1-|\Gamma_L|^2}{|1-\Gamma_{OUT}\Gamma_L|^2} \tag{1.4}$$

where

$$\Gamma_{OUT} = s_{22} + \frac{s_{12}s_{21}\Gamma_S}{1-s_{11}\Gamma_S} \tag{1.5}$$

In the above expressions Γ_{IN} represents the true input reflection coefficients of the two-port, with an arbitrary load termination, Γ_L. Similarly, Γ_{OUT} stands for the output reflection coefficient of the two-port, with an arbitrary source termination connected to the input. It should be clear that when there is no interaction between the input and output ports ($|s_{12}| = 0$), then (1.3) and (1.5) are simplified to

$$\Gamma_{IN} = s_{11}$$

and

$$\Gamma_{OUT} = s_{22}$$

Viewing (1.2) and (1.4) gives a somewhat easier interpretation of the composition of the overall gain, since both of them have three distinct portions. For example, (1.2) could be expressed as

$$G_T = G_{1D}(G_0)G_2 \tag{1.6}$$

where

$$G_0 = |s_{21}|^2 \tag{1.7}$$

is the basic Z_0-based transducer power gain, and

$$G_{1D} = \frac{1-|\Gamma_S|^2}{|1-\Gamma_{IN}\Gamma_S|^2} \tag{1.8}$$

is the transducer gain-factor change due to the selection of Γ_S and Γ_L. Although Γ_L is not shown directly in (1.8), remember that Γ_{IN} is a function of all four *S*-parameters and Γ_L. The third term of (1.6),

$$G_2 = \frac{1-|\Gamma_L|^2}{|1-s_{22}\Gamma_L|^2} \tag{1.9}$$

indicates the change of the transducer gain due to the load selection, Γ_L.

We can summarize (1.8) and (1.9) by stating that when the load is changed from Z_0 to any arbitrary value, there is both a direct and an indirect effect on the transducer gain expression of (1.2). When the source impedance is changed from Z_0 to an arbitrary value, it also has a direct and indirect effect on G_T.

When the source and load terminations are both equal to Z_0, then

$$\Gamma_S = \Gamma_L = 0$$

In that case, both (1.2) and (1.4) are reduced to

$$G_T = G_0 = |s_{21}|^2$$

Finding the transducer power gain requires knowledge of the *S*-parameters, as well as the source and load terminations connected to the two-port. During linear circuit simulation, the source and load terminations are either given or computed from the circuit topology description. The two-port's *S*-parameters are either specified or computed from a linear device model.

The *S*-parameter technique for designing amplifiers is based on finding the two terminations of a two-port to provide the desired gain. However, that requires solving (1.1) with two unknowns, Γ_S and Γ_L, which cannot be done. One exception is when the goal is the maximum gain of the two-

port ($G_T = G_{MAX}$). In that case, we can write two expressions with two unknowns—for which we can find a solution. This approach, shown in Chapter 2, is very useful when designing amplifiers for maximum small-signal power gain, G_{MAX}, with simultaneously conjugate matched terminations.

The transducer gain expression is not useful to design amplifiers for an arbitrary gain less than the maximum gain. That is, for $G_T < G_{MAX}$ there is no practical procedure that can be used to determine the source and load terminations for a given G_T, except trial and error. For such cases, we, however, will develop two additional gain definitions, called *available power gain* and *operating power gain* [7]. As we will see in Chapter 2, these additional techniques are very useful in *low-noise* and *linear power* amplifier design, where one of the two ports of the active device is tuned for special performance, other than maximum gain.

When a signal applied to the input of a two-port produces a response at the output, but a signal applied to the output has no effect on the input, the two-port is unilateral. In a bilateral two-port, signals flow in both directions. If the input and output ports of our active devices were perfectly isolated from each other, the transducer gain approach would offer a straightforward approach to small-signal amplifier design. Unfortunately, all physical transistors have internal feedback elements, such as Miller capacitance and package parasitics, that make the device bilateral. There are three ways to handle this undesirable input-output interaction: (1) tune it out, (2) pretend it does not exist, or (3) deal with it mathematically. The first approach is called neutralization [8], and the second one is generally

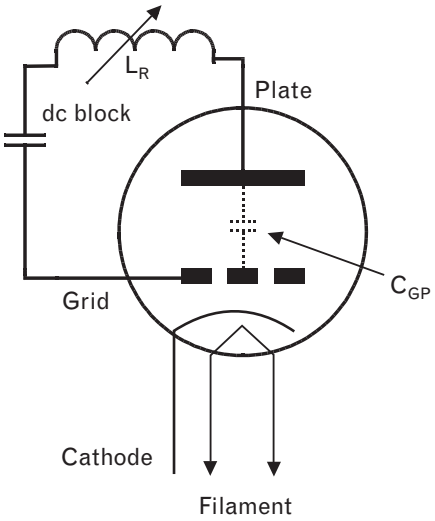


FIGURE 1.3
Neutralizing a vacuum tube's grid-to-plate capacitance, C_{GP}, with an external dc blocked parallel resonant feedback inductor, L_R.

4. Strictly speaking, a unilateral two-port has signal transmission only one direction, so a neutralized device is also unilateral. Still, the above-mentioned names are what most people in the industry use.

referred to as the unilateral design.[4] The third technique, called bilateral design, is covered in Chapter 2.

1.3 Neutralization

The original concept of neutralization dates back to the days when electron tubes were used in RF circuits where interelectrode capacitances formed undesirable feedback. Variations in the load connected to the output of a vacuum tube caused pulling effects on the tuning circuit at the input side and changed the circuit's performance. The undesirable feedback could be eliminated by placing an external tuning inductor between the input and output (Figure 1.3), creating a resonant circuit. Since the internal capacitance was lossless, it was possible to resonate it at a single frequency to improve the isolation of the tube. By eliminating the internal feedback, we can set to nearly $|s_{12}| = 0$. However, the other three *S*-parameters are also changed by neutralization. Although parallel resonance applied only to a single frequency, the performance improvement of the tube was noticeable through a 5% to 10% bandwidth.

Applying neutralization to solid-state devices is more difficult because the primary internal feedback mechanism, such as the Miller capacitance in a bipolar transistor, is no longer a lossless component. In addition, since transistors have lower impedances than tubes, the series common-lead inductance also forms feedback that needs to be accounted for [Figure 1.4(a)]. In such cases a second feedback circuit is necessary. By also adding

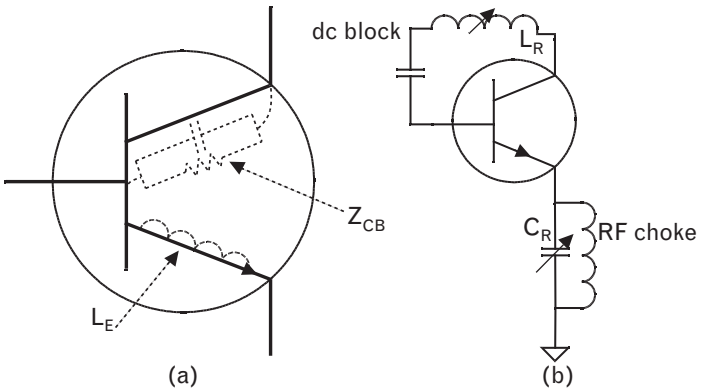


FIGURE 1.4 (A) *Two important feedback components of a bipolar transistor: lossy collector-base capacitance, Z_{CB}, and lossy emitter inductance, L_E.* (B) *Dual neutralization by adding parallel and series feedback sections, L_R and C_R. Since the series feedback must be bypassed for the dc current, it is not a very practical solution.*

5. An inductor with very high RF impedance and low dc resistance.

external series capacitive feedback, we can always obtain perfect isolation [Figure 1.4(b)]. However, series feedback is harder to realize and requires additional dc bypass circuitry, such as an RF choke.[5] For practical considerations, the parallel inductive feedback alone brings a significant improvement, by reducing $|s_{12}|$ while increasing $|s_{21}|$. Eliminating the internal parallel feedback generally also increases input and output impedances.

If neutralization is so great, we may wonder why device manufacturers do not apply it to their products. There are good reasons for that. First, neutralization is a narrowband solution, applicable to selected frequencies only. Suppliers would need to tune devices for the various commonly used frequency bands. Second, the RF feedback elements may not fit into the RF package, and they must also be combined with dc bias blocking/bypassing circuitry. Finally, and *very importantly,* the added external resonator element(s) may represent *positive feedback* at some other frequencies—possibly leading to oscillation.

To summarize, partial or full neutralization offers improved input-output isolation and higher gain at the cost of increased circuit complexity and possible RF stability problems. It may also require additional dc bias elements, like the RF choke. Therefore, neutralization must always be customized by the end user for specific applications. We must emphasize again, however, that while the added neutralization network effectively tunes out the input-output interaction for a narrow frequency range, it may cause problems at other frequencies. Therefore, the complete circuit must be very carefully analyzed for broadband RF stability—a subject we will cover in Section 1.5.

1.4 Unilateral transducer gain

An approximate design procedure, called the *unilateral design*, simply ignores the interaction between the input and output ports. In this case we assume that the input and output reflection coefficients of the device are always equal to the original measured s_{11} and s_{22} in a 50-Ω system, regardless of the actual terminations connected to the device in the final circuit.

Before we proceed, let us summarize the fundamental difference between neutralization and the unilateral technique, because they are both based on $|s_{12}| = 0$. In the former we modify the other two-port *S*-parameters to show the effect of neutralization. Therefore, it is an *exact procedure*. In contrast, the unilateral approach simply sets the reverse transmission parameter magnitude to zero and leaves the other three *S*-parameters unchanged. As a result, a unilateral design is only an *approximate* technique that may or may not lead to acceptable performance.

If we simply set $|s_{12}| = 0$ in (1.1), the *transducer gain* expression is simplified to the *unilateral transducer gain*,

$$G_{TU} = \frac{1-|\Gamma_S|^2}{|1-s_{11}\Gamma_S|^2}|s_{21}|^2 \frac{1-|\Gamma_L|^2}{|1-s_{22}\Gamma_L|^2} = G_{1U}(G_0)G_{2U} \tag{1.10}$$

where G_{1U}, G_0, and G_{2U} are the three independent components of the total unilateral gain. In the unilateral approach, since there is perfect isolation between the input and output ports, the terminations we choose have no effect on each other.

The maximum unilateral transducer gain is achieved when Γ_S and Γ_L are set to s_{11}^* and s_{22}^*, respectively. Under these conditions the maximum value of unilateral gain is obtained, namely

$$G_{TUMAX} = G_{1UMAX}G_0G_{2UMAX} = \frac{1}{(1-|s_{11}|^2)}|s_{21}|^2 \frac{1}{(1-|s_{22}|^2)} \tag{1.11}$$

Looking at the two fractional portions of (1.11), we should recognize that they represent the reciprocals of mismatch losses. Since the expression

$$ML_1 = 1-|s_{11}|^2$$

represents the mismatch loss (defined in Volume I, Chapter 2) between a Z_0 source and a device having an input reflection coefficient s_{11}, the inverse of that quantity

$$\frac{1}{ML_1} = \frac{1}{(1-|s_{11}|^2)} = G_{1UMAX} \tag{1.12}$$

is the portion of the power gain realized when the mismatch is eliminated at the input port. The same applies for the output side of the two-port:

$$\frac{1}{ML_2} = \frac{1}{(1-|s_{22}|^2)} = G_{2UMAX} \tag{1.13}$$

Equations (1.12) and (1.13) indicate the exact maximum gain increase at the input or output ports, respectively, for a truly neutralized device, without interaction between the two matching networks. When $|s_{12}| \neq 0$, the gain computed by (1.11) is not exact. Since there is an ambiguity, let us now find the magnitude of the error involved with the unilateral assumption.

1.4.1 Unilateral figure of merit

A common (and erroneous) belief is that a low $|s_{12}|$ alone always leads to negligible input-output interaction, and such a decision can lead to significant errors. For example, even though a common-base configuration of a bipolar transistor has very good isolation between its two ports, changing one of the terminations significantly affects the impedance of the adjacent port. The reason for this strong interaction is that the output and input signals are nearly in-phase, leading to *positive* feedback through any internal output-input coupling.

An alternative and more reliable way to evaluate the effect of input-output interaction is to compute the *unilateral figure of merit* [7], sometimes called the *U*-factor, which is a function of *S*-parameters and is therefore frequency dependent:

$$U = \frac{|s_{11}s_{22}s_{12}s_{21}|}{(1-|s_{11}|^2)(1-|s_{22}|^2)} \tag{1.14}$$

Examining the above expression reveals that while U is directly proportional to the magnitudes of all four two-port *S*-parameters, it may be more seriously affected by reflection coefficients if their magnitudes are close to unity. The *U*-factor is very helpful in estimating the difference between the computed unilateral gain and the actual gain by setting limits for the maximum error.

If we set up a ratio of (1.1) and (1.10), (G_T/G_{TU}), and convert it to decibels, we can bound the error caused by the unilateral assumption:

$$10\log\left[\frac{1}{(1+U)^2}\right] < 10\log\left[\frac{G_T}{G_{TU}}\right] < 10\log\left[\frac{1}{(1-U)^2}\right] \tag{1.15}$$

which is interpreted as

Maximum negative decibel error < Actual decibel error < Maximum positive decibel error

where the largest possible absolute error is

$$\text{decibel error}_{MAX} = \text{Positive decibel error} + |\text{Negative decibel error}|$$

If the computed value of the G_T/G_{TU} ratio in (1.15) is close to unity, the device is a good candidate for the unilateral approach. Table 1.1 and Figure 1.5 compare the maximum possible decibel error associated with various U values, showing that as U approaches unity the gain error

TABLE 1.1 MAXIMUM POSSIBLE POSITIVE AND NEGATIVE ERRORS CAUSED BY THE UNILATERAL ASSUMPTION VERSUS THE *U*-FACTOR

U	− ERROR (dB)	+ ERROR (dB)
0.01	0.09	0.09
0.02	0.17	0.18
0.05	0.42	0.45
0.1	0.83	0.91
0.2	1.58	1.94
0.5	3.52	6.02
0.9	5.58	20.00
0.95	5.80	26.02

Note: A computed value of $U = 0.2$ means that the actual gain, G_T, of a two-port may be 1.58 dB lower or 1.94 dB higher than the computed unilateral gain, G_{TU}.

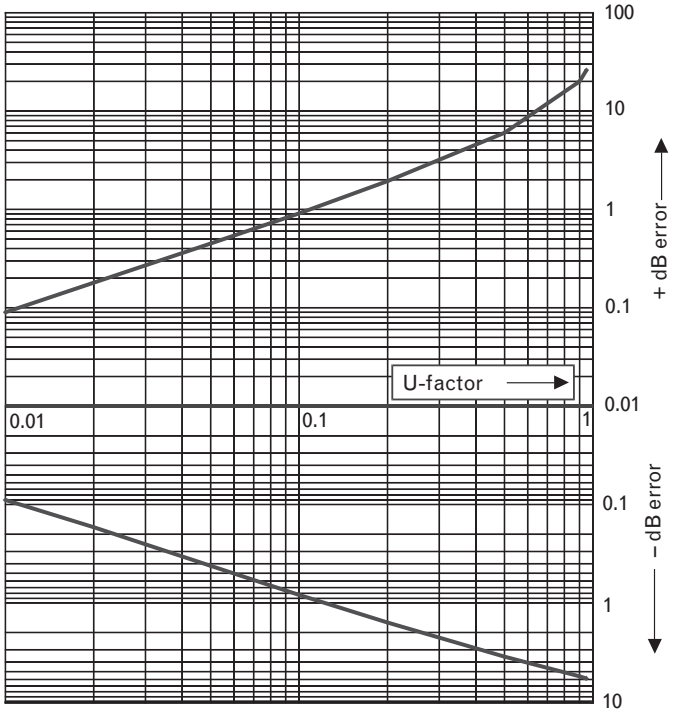


FIGURE 1.5
Plotted version of Table 1.1 shows how quickly the error range increases for larger U-factors.

becomes excessive. A low *U*-factor leads to a small error when using unilateral design. As a general rule, devices with *U* less than 0.1 bring less than ±1-dB error. For higher *U* values, the unilateral assumption is not recommended.

1.4.2 Illustrative example: unilateral gain calculations

Compute the highest and lowest gains by using the unilateral assumption for an Infineon BFP640 bipolar device. Measured *S*-parameters (2V, 20 mA) at 900 MHz are given as:

$s_{11} = 0.40 \angle -102°$;

$s_{21} = 20.7 \angle 106°$;

$s_{12} = 0.029 \angle 60°$;

$s_{22} = 0.54 \angle -43°$.

The basic 50-Ω transducer gain is

$$G_{0\,dB} = 10\log(|s_{21}|^2) = 10\log(428.5) = 26.3\,\text{dB}$$

The maximum gain improvement at the input and output ports are

$$G_{1UMAXdB} = 10\log\left[\frac{1}{(1-|s_{11}|)^2}\right] = 10\log\left[\frac{1}{(1-|0.4|^2)}\right] = 0.76\,\text{dB}$$

$$G_{2UMAXdB} = 10\log\left[\frac{1}{(1-|s_{22}|)^2}\right] = 10\log\left[\frac{1}{(1-|0.54|)^2}\right] = 1.5\,\text{dB}$$

Therefore, the maximum unilateral gain in decibels is

$$G_{TUMAXdB} = G_{1UMAXdB} + G_{0dB} + G_{2UMAXdB} = (0.76 + 26.3 + 1.5)\text{dB} = 28.56\ \text{dB}$$

Use (1.14) to calculate the *U*-factor from the given *S*-parameters.

$$U = \frac{|s_{11}||s_{22}||s_{12}||s_{21}|}{(1-|s_{11}|^2)(1-|s_{22}|^2)} = \frac{(0.4)(0.54)(0.029)(20.7)}{(1-0.4^2)(1-0.54^2)} = 0.22$$

Then, from (1.15), we can calculate the maximum error range:

$$\text{Negative error limit} = 10\log\left[\frac{1}{(1+U)^2}\right] =$$

$$10\log\left[\frac{1}{(1+0.22)^2}\right] = -1.73\,\text{dB}$$

$$\text{Positive error limit} = 10\log\left[\frac{1}{(1-U)^2}\right] =$$

$$10\log\left[\frac{1}{(1-0.22)^2}\right] = 2.16\,\text{dB}$$

Designing an amplifier for maximum gain under the unilateral assumption, the actual gain in decibels may be anywhere between

$$G_{TLOWdB} = G_{TUMAXdB} - |\text{Negative error limit}| = (28.56 - 1.73) = 26.83 \text{ dB}$$

and

$$G_{THIGHdB} = G_{TUMAXdB} + \text{Positive error limit} = (28.56 + 2.16) = 30.72 \text{ dB}$$

which is nearly a 4-dB error range. Clearly, an amplifier design should not be based on such large possible error.

Although the unilateral design is only an approximation, it is often used as a quick estimate or to provide initial values for circuit optimization. However, it is not recommended when the value of the *U*-factor exceeds 0.1. Bilateral design ($|s_{12}| \neq 0$) eliminates errors and should always be used. Furthermore, the bilateral design procedure is very powerful when used with appropriate CAD tools.

1.4.3 Amplifier design with single matching networks

A special case exists between the unilateral and bilateral methods where s_{12} does not affect the accuracy of the computed gain. If the performance requirements are such that they can be fulfilled by using only an input or output network, the computed gain is exact. Although in most applications both input and output matching circuits are used, we want to illustrate the single matching case also because it will help us later to better understand the uses of constant-gain circles [7].

1.4.3.1 Matching network added to the output port only

When we use a Z_0 source ($\Gamma_S = 0$) and an output matching network to transform the load to an arbitrary value Γ_L, as shown in Figure 1.6, the transducer power gain expression is simplified to

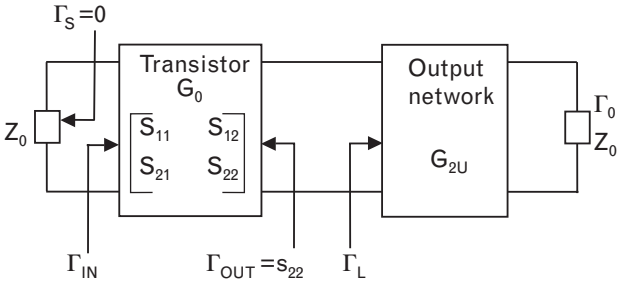


FIGURE 1.6
Amplifier with single matching network at the output. In this case the exact transducer gain is a function of s_{21}, s_{22}, and the load connected to the two-port, Γ_L.

$$G_T = |s_{21}|^2 \frac{1-|\Gamma_L|^2}{|1-s_{22}\Gamma_L|^2} = (G_0)(G_{2U}) \tag{1.16}$$

where G_0 and G_{2U} represent the basic 50-Ω transducer gain and the change in gain due to an arbitrary load selection.

Examining (1.16) shows that the gain is only a function of two *S*-parameters, s_{21} and s_{22}, and the load termination connected to the two-port. We can see that with given *S*-parameters, the load selection controls the gain. More specifically, the relationship of Γ_L with respect to s_{22} is very important. Later we will also solve the expression for Γ_L to find out what kind of load is needed for a specified gain. Before we do that, however, let us look at the other single-side matching case.

1.4.3.2 Matching network added to the input port only

When a two-port is driven by an input network presenting an arbitrary source Γ_S and a Z_0 load is used (i.e., $\Gamma_L = 0$) as shown in Figure 1.7, the transducer power gain is reduced to

$$G_T = \frac{1-|\Gamma_S|^2}{|1-s_{11}\Gamma_S|^2}|s_{21}|^2 = (G_{1U})(G_0) \tag{1.17}$$

Once again, the transducer gain is only a function of two *S*-parameters and one termination. Equation (1.17) is exact, and it may be solved for the source termination as a function of the *S*-parameters and a specified gain.

Most practical amplifiers have both input and output matching networks. Applying only a single network may not give us the desired gain and may also lead to a poor impedance match at the port adjacent to the matching network. We mentioned earlier, however, that the terminations have a predictable effect on the gain, and we will now look at some graphical tools that may be very helpful to find the optimum terminations.

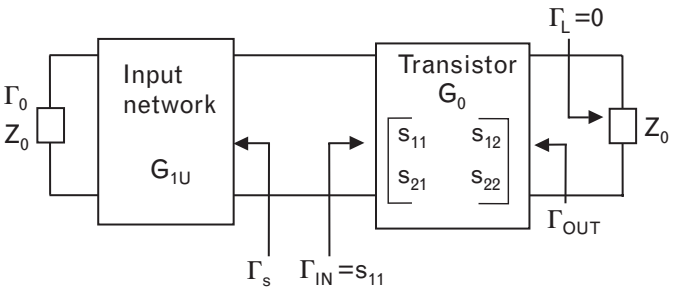


FIGURE 1.7
Amplifier with single matching network at the input. In this case the exact transducer gain is a function of s_{21}, s_{11}, and the source connected to the two-port, Γ_S.

1.4.4 Unilateral constant gain circles

In (1.10), (1.16), and (1.17), G_{1U} and G_{2U} [rewritten here for convenience as (1.18) and (1.19)] represent the terms in the transducer power gain relation that are a function of an input or output port reflection coefficient (i.e., of the termination connected to that port). Stating it another way, when the source or load termination is changed from Z_0 to an arbitrary value, the overall gain also changes. When Γ_S has zero magnitude in (1.18), the value of G_{1U} is unity, or 0 dB. Similarly, when Γ_L has zero magnitude in (1.19), the value of G_{2U} is also unity. However, there are other source and load terminations that will also lead to 0-dB change in gain. Keep in mind also that the magnitudes of G_{1U} and G_{2U} may increase above 1.0, meaning that we improved the match at that port, or become less than 1.0, meaning we made the impedance match even worse.

$$G_{1U} = \frac{1-|\Gamma_S|^2}{|1-s_{11}\Gamma_S|^2} \tag{1.18}$$

$$G_{2U} = \frac{1-|\Gamma_L|^2}{|1-s_{22}\Gamma_L|^2} \tag{1.19}$$

The highest and lowest values for G_{1U} and G_{2U} are

$$0 \le G_{1U} \le G_{1UMAX}$$

and

$$0 \le G_{2U} \le G_{2UMAX}$$

where a zero magnitude means total mismatch, such as an open or short circuit. G_{1UMAX} and G_{2UMAX} represent the best we can get by conjugate matching the input or output port, respectively, as shown in (1.12) and (1.13).

To find out what kind of terminations lead to a specific amount of change, we can solve (1.18) and (1.19) for Γ_S and Γ_L, respectively. The solutions come in the form of circle equations. For both source and load terminations we can now plot a family of constant gain circles, as shown in Figure 1.8. The gain circles are referenced to a unit-radius Smith chart.[6]

1.4.5 Illustrative example: single-sided amplifier design

Task: The Infineon BFP 405 device has a basic gain of 14.7 dB at 1,900 MHz in a 50-Ω system. Use the unilateral constant-gain circles to find a

6. Unless otherwise specified, we always refer to a normalized unit-radius Smith chart, having a radius of 1.0. Details of the chart are covered in Volume I, Chapter 4.

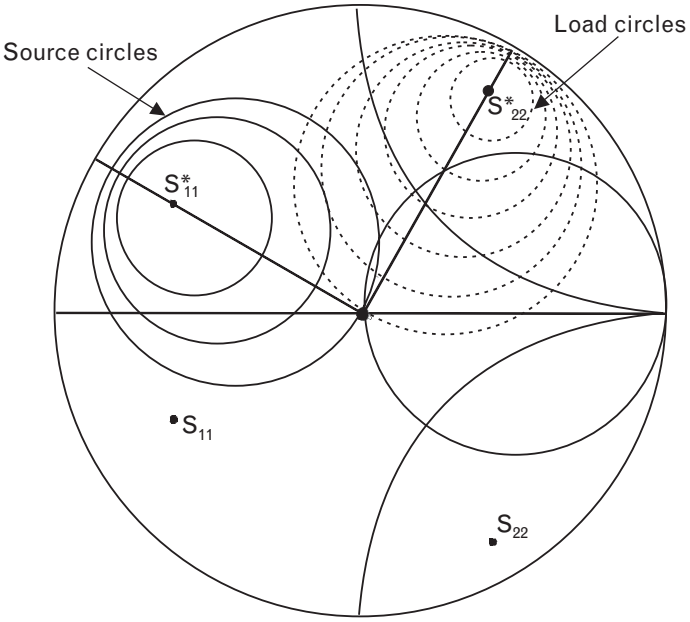


FIGURE 1.8
Unilateral constant-gain circles plotted on the Smith chart. Each circle represents the locus of source or load terminations that cause a specific gain change.

matching network that increases the gain by 2.3 dB, to 17 dB. Device *S*-parameters, measured at 2V, 2 mA, are given in Table 1.2.

Solution: First we need to find if there is at least 2.3-dB mismatch at the input or output ports. If yes, we need to select the port with which we want to work. The decibel mismatch loss at the input port, computed from $|s_{11}|$, is

$$\begin{aligned} ML_1 &= 10(\log(1-|s_{11}|^2)) \\ &= 10(\log(1-(0.707)^2)) \\ &= 3 \text{ dB} \end{aligned}$$

Since the mismatch loss is greater than the 2.3 dB we need, a matching network can easily be found by graphical Smith chart techniques. For illustrative purposes, we plotted several of the constant-gain source circles in Figure 1.9 to show that the basic 50-Ω gain of the device may be either increased or reduced by choosing the appropriate termination. For gain higher than $10(\log|s_{21}|^2)$, we choose a termination on a positive decibel

TABLE 1.2 TABULATED TWO-PORT 50-Ω *S*-PARAMETERS OF THE BFP 405 DEVICE, MEASURED AT 1,900 MHz

s_{11}	s_{21}	s_{12}	s_{22}
$0.707 \angle -67°$	$5.45 \angle 119°$	$0.058 \angle 55°$	$0.84 \angle -32°$

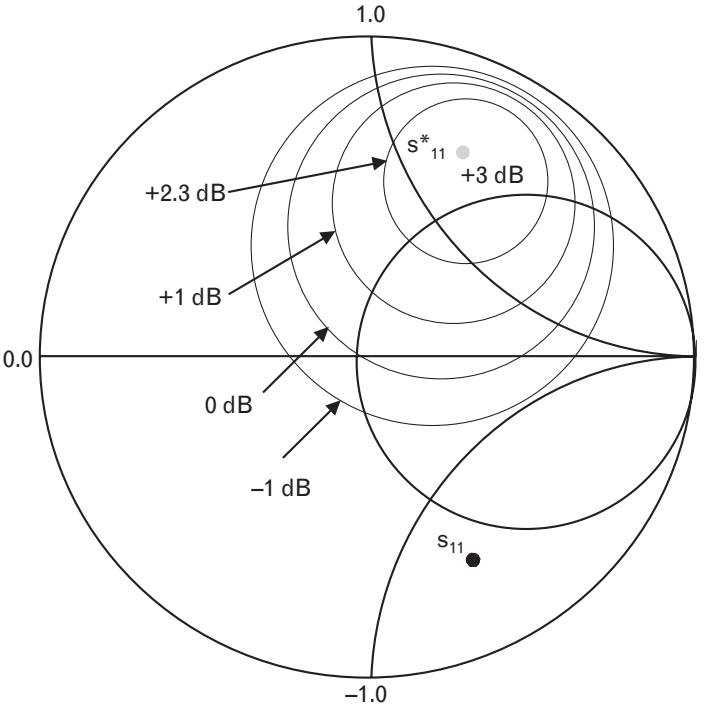


FIGURE 1.9
Five constant-gain source circles, ranging from +3 dB to −1 dB, for the BFP 405 device at 1,900 MHz. Note that the 3-dB constant-gain circle is just a single point (circle with zero radius), located at the complex conjugate of s_{22}. The 0-dB circle must always pass through the center of the Smith chart.

gain circle. For example, any source selected from the 2.3-dB constant gain circle increases the gain by 2.3 dB. To lower the gain, we apply a termination located on a negative decibel circle.

The mismatch loss at the output port is

$$\begin{aligned} ML_2 &= 10(\log(1-|s_{22}|^2)) \\ &= 10(\log(1-(0.84)^2)) \\ &= 5.3\,\text{dB} \end{aligned}$$

Since the mismatch losses at both ports exceed 2.3 dB, for RF gain considerations we could apply a matching network at either side, as shown in Figure 1.10.

Plotting the two 2.3-dB constant-gain circles of the source and load sides on two separate Smith charts (Figure 1.11) helps us to select the necessary circuit topologies. If we start from 50Ω, the center of the chart, a series C–parallel L highpass network combination can transform the impedance to either side of the gain-circles and offer two possible solutions for both sides. We will simulate the frequency response of this amplifier with all four combinations of single-sided matching networks to show that all of them provide exactly 17-dB gain at 1,900 MHz. Then, we will also show what happens when we apply an appropriate network to both sides.

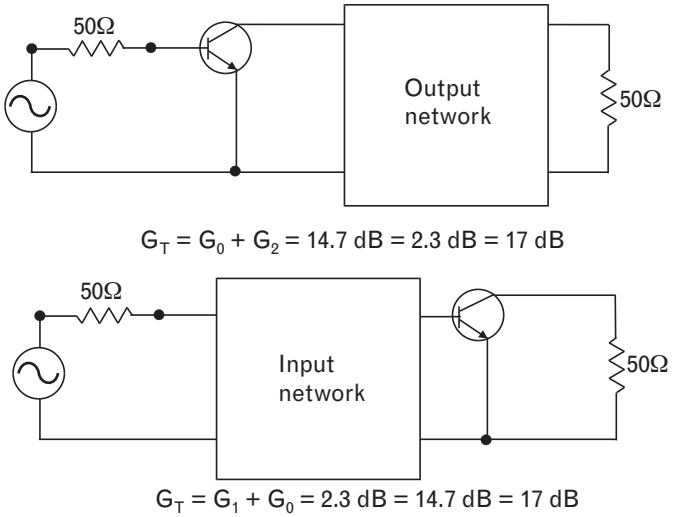


FIGURE 1.10
To have 17-dB total gain, we may add a network to either side of the active device. Since a passive network does not provide gain, the matching network must reduce the existing mismatch loss of the port by 2.3 dB. Figure 1.11 shows how the circuits are determined.

All four matching sections of Figure 1.11 are in highpass configurations; therefore, their frequency responses roll off at low frequencies. Since the fundamental gain response of the transistor rolls off at the high frequencies, it is a good idea to select highpass circuit topologies for the matching section to obtain a more balanced bandpass response. Highpass matching topologies are also convenient for dc biasing, as we will see in Section 1.8.3. Collector and base currents may be passed through parallel inductors, while series capacitors may serve as dc blocking elements.

Assuming lossless matching elements at the 1,900-MHz design frequency, we should get exactly the expected 17-dB gain as long as we only add one network to the device. It makes no difference whether we use a network on the input or the output side. The total gain is the basic 14.7 dB of the device plus 2.3 dB recovered from the existing mismatch of one of the ports.

If we were to have a truly unilateral device here, the gain would increase by another 2.3 dB to 19.3 dB when we also apply one of the output matching networks to the adjacent port. However, for the real-life device ($|s_{12}| \neq 0$) the true gain can be *greater or less* than 19.3 dB, due to the input-output interaction. Using the *U*-factor allows us to compute the amount of uncertainty.

Figure 1.12 shows the frequency responses when using a single matching network, and also with a network added to both sides. The results of the single-sided matching exactly follow our expectations—2.3 dB more than the basic gain of the device. However, the two-sided match with our unilateral approach confirms the effect of input-output interaction. The gain actually *decreased* from 17 dB when the second matching section is added.

Frequency response with the highpass type input matching networks added shows gain roll-off at the lower frequencies. The same effect is not noticeable with Out2 output network and with the two-sided match

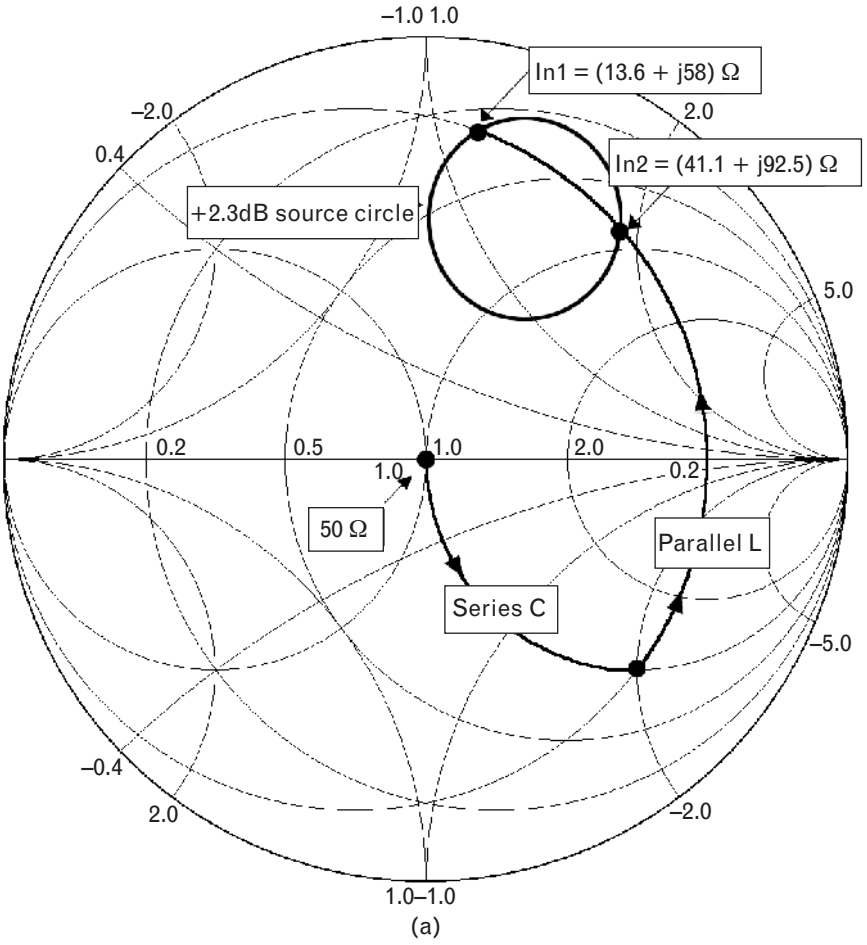


FIGURE 1.11
Immitance paths of the two-element highpass matching networks plotted on a ZY Smith chart. By selecting different values for the parallel inductor, we can realize (a) two input networks (moving to either In1 or In2) and (b) two output networks (moving to either Out1 or Out2). All four circuit options provide 2-dB gain increase.

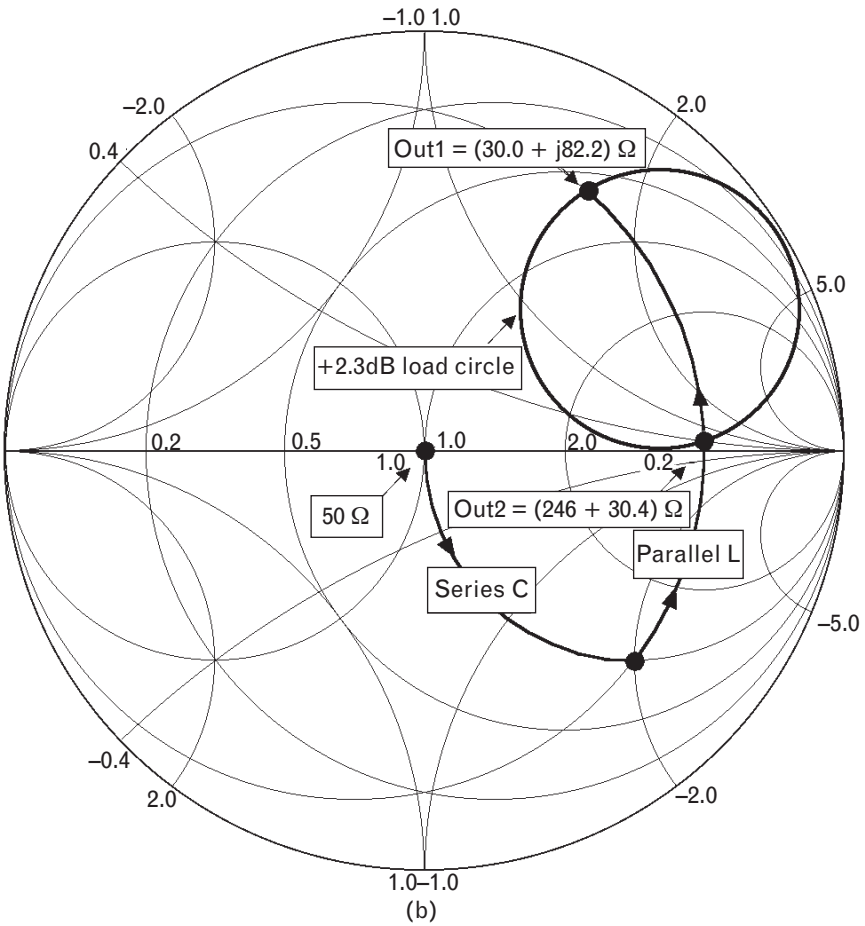
through this frequency range. The roll-offs, however, are visible when the simulation is extended to lower frequencies.

Bilateral design (covered in Chapter 2) eliminates the error caused by the unilateral approach. However, it requires a little more analytical work and also an investigation for RF stability, as shown in the next section.

1.5 RF circuit stability considerations

One of the most frustrating experiences of RF engineering is when a newly designed amplifier oscillates instead of functioning as intended. Virtually every designer in that field, including us authors, had such an unpleasant experience. The oscillation may be "fixed" by shielding, tweaking, painting, dampening the circuit, but at times, even a short period of oscillation may cause permanent damage.

FIGURE 1.11
Continued.



Experiencing unwanted oscillation in the design laboratory is exasperating, but once the cause is found the fix is generally simple. When it shows up on the production line, it can cause lengthy and expensive production delays. Oscillation at the customer's site is very expensive to fix, and that customer may never buy your product in the future.

A circuit is unstable when a signal can increases without any limit. Actually, nonlinearities do limit the maximum signal level and either set it into a steady-state oscillation or stop it completely. In Chapter 6 we provide a more detailed explanation of why and how oscillators work, and in this chapter we only examine the potential to start oscillation and how to prevent it.

If stability is so important, why don't we get a warning from the device manufacturers about possible RF instability? The easy answer is easy—virtually all RF/MW transistors are potentially unstable at some frequencies, and sellers do not want to advertise potential problems to those not wanting oscillation. They may warn you about grounding and coupling concerns, but do not expect their datasheets to list the stability-factor, which is analytically defined in the next section.

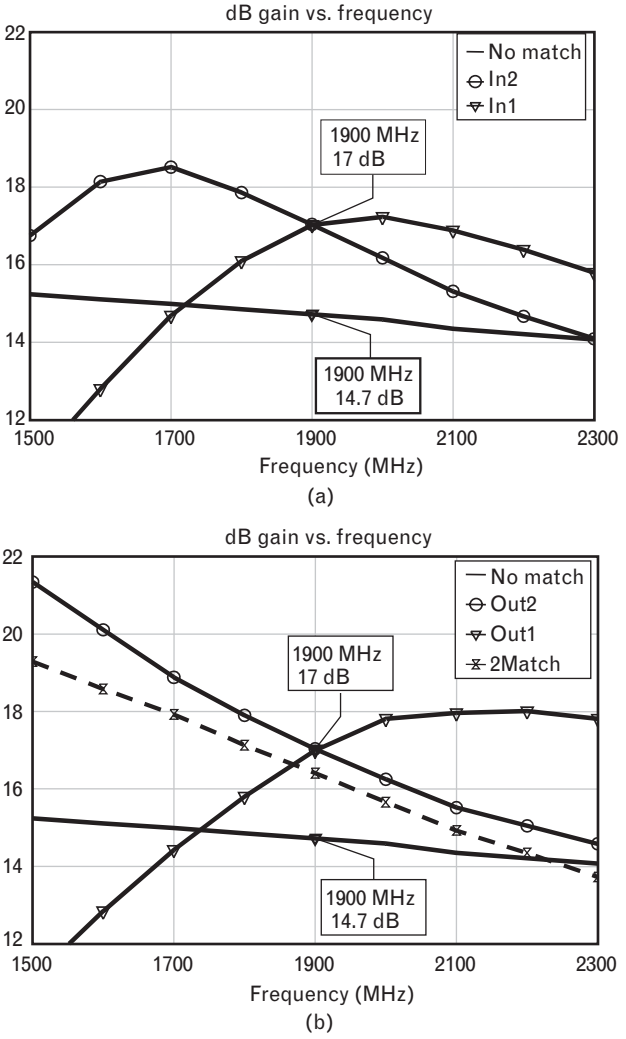


FIGURE 1.12
Comparison of the four single-sided amplifier matching illustrations derived from the Smith chart of Figure 1.11. (a) Labels In1 and In2 refer to input matching networks only, while (b) Out1 and Out2 are with output matching only (see Figure 1.10). "No match" shows the basic 50-Ω gain, and "2Match" stands for adding In2 and Out2 simultaneously to both sides.

In low-frequency analog circuits, where transfer functions are commonly available, the Nyquist criteria [9] provide a safe indication of stability. At RF and microwave frequencies circuit and system designers face a much more difficult and tedious task because transfer functions are virtually never given in closed form.[7] Therefore, a thorough stability analysis should be performed through a wide range of frequencies, input signal levels, and external terminations.

Since true broadband nonlinear models are not always available for the active devices, RF circuit stability is most conveniently evaluated at individual frequencies, based on small-signal two-port *S*-parameters. Such an approach is sufficient only for linear, small-signal circuit applications,

7. Some of the commercially available RF circuit/system simulators provide the Nyquist test in simple and convenient form.

although some of the concepts used here may also be useful as a start for large-signal analysis. Later in this book we will also look at stability analysis under large-signal (nonlinear) operation.

RF/MW circuit and system engineers are becoming more aware of stability-related problems and are more willing to spend time on stability analysis. A common mistake, however, is to examine only through the passband of the system, which is not sufficient. When out-of-band instability is neglected, it may lead to unwanted low-frequency or high-frequency oscillation.

I, Les Besser, once witnessed an intended 8- to 12-GHz balanced amplifier behaving more like a 50-MHz comb-generator.[8] The frustrated design engineer showed me his stability analysis performed between 2 and 20 GHz, predicting unconditional stability. Asking why he did not cover the lower frequencies, he replied, "I did not have data below 2 GHz. Also, I figured that even if the device would oscillate at low frequency, those signals could not pass through the directional couplers of the amplifier."

We definitely do not want our amplifiers to oscillate at any frequency, for several reasons. Let us look at a few of them:

- When oscillation takes place, the active device is pushed into its large-signal mode and the performance changes very significantly. The small-signal *S*-parameters are no longer valid, and therefore, the circuit design is incorrect.
- When a device oscillates it becomes more noisy.
- Even if the oscillation is far below the passband of the amplifier, as was the case above, the newly created signal mixes with any incoming signal and shows up at the output.
- Oscillation may damage the active device(s).

Now that we have gotten your attention, let us find out what causes oscillation and how to avoid it in amplifier design.

1.5.1 What may cause RF oscillation

One fundamental approach to create oscillation is to feed part of the output signal of an amplifier back to its input in such a way that the phase angles of the two signals are exactly the same (Figure 1.13). If some other nonlinear conditions are also satisfied, steady-state oscillation develops, as shown in Chapter 6. Oscillator designers intentionally use positive feedback because potential oscillation exists through a wide range of frequencies, up to f_{MAX} of the device (f_{MAX} is where the matched neutralized gain of the two-port

8. A signal source that provides a large number of harmonically related frequencies of nearly equal magnitudes.

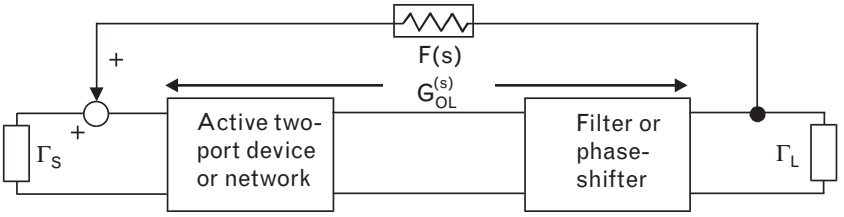


FIGURE 1.13 *Simplified block diagram of a phase-shift oscillator. The output of the active device is filtered and the phase is adjusted until the input and feedback signals, summed at the input port, are in the same phase. In real physical circuits and systems, unintentional feedback paths can perform the same functions, leading to parasitic oscillation.*

drops to unity). Therefore, one possible way to prevent unwanted oscillation is to eliminate undesirable feedback in the amplifier's circuitry.

The closed-loop gain of the feedback circuit, $G_{CL}(s)$, is given by

$$G_{CL}(s) = \frac{G_{OL}(s)}{1 - G_{OL}(s)F(s)} \tag{1.20}$$

where $G_{OL}(s)$ is the open-loop gain of the amplifier-filter combination, including nonlinear and limiting effects, $F(s)$ is the feedback path gain, $G_{OL}(s)F(s)$ is the loop gain, and $s = j\omega$ is a complex variable—*no connection to scattering parameters*.

Oscillation takes place when the denominator of (1.20) reaches zero value.

Unfortunately for amplifier designers, components or circuitry causing harmful feedback are often not obvious; instead they come in many unexpected ways. Unplanned positive feedback may be caused by poor component grounding, inductive or capacitive coupling, waveguiding effects, and even by poorly filtered dc bias networks. Many times even the most sophisticated state-of-the-art circuit simulators cannot detect the presence of the problem, and we need to use other tools, such as electromagnetic simulators, to find the right solution. Even though such investigations may be tedious and time-consuming, they may be necessary to prevent unwanted oscillation. Therefore, it is a wise and recommended investment of one's time.

What is the frequency range through which an active device can create desired or undesired oscillation? The feedback oscillator of Figure 1.13 may be designed for any frequency, as long as the active device has more than unity gain. The low-frequency limit of such oscillation is only controlled by the coupling elements. At the high end of the spectrum, since the gain rolls off, the limiting frequency is where the highest achievable (i.e., neutralized) gain, U_p, drops to unity (Figure 1.14). This neutralized

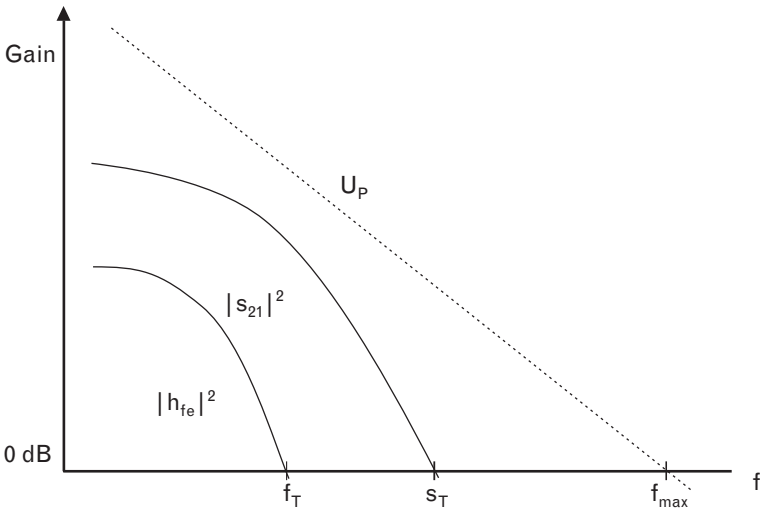


FIGURE 1.14
U_P, the computed maximum neutralized gain, is a theoretical figure. It assumes that a two-port is first neutralized and then matched at both ports. It should not be confused with G_{UMAX}, which is computed under the unilateral assumption, simply setting $|s_{12}|$ to zero.

maximum gain [10] can be expressed in terms of the two-port *S*-parameters, as

$$U_P = \frac{\left|\frac{S_{21}}{S_{12}} - 1\right|^2}{2K\left|\frac{S_{21}}{S_{12}}\right| - 2\,\mathrm{Re}\left(\frac{S_{21}}{S_{12}}\right)} \tag{1.21}$$

where *K*, commonly called the stability *K*-factor, is a frequency-dependent function of the two-port *S*-parameters. We show it here for convenience and will further deal with it in Section 1.5.3.1.

$$K = \frac{1 - |s_{11}|^2 - |s_{22}|^2 + |s_{11}s_{22} - s_{12}s_{21}|^2}{2|s_{12}s_{21}|} \tag{1.22}$$

Since feedback-type oscillation requires greater-than-unity gain from the active device, in addition to f_{MAX} being the highest theoretical frequency for amplification, it is also the limit for oscillation; f_T, the frequency where the computed short-circuited current-gain of a bipolar transistor drops to unity value, is generally not used in *S*-parameter design other than for comparing it with other devices.

Summing up our introductory stability discussion, we must accept that up to f_{MAX} even the most carefully designed amplifier may oscillate due to external RF feedback paths. Of course, one may argue that we are describing the worst-case condition since the device is generally not neutralized. For a device that is not neutralized, the simultaneously matched device gain will drop below unity magnitude around 80% to 90% of f_{MAX}. Still, since f_{MAX} of a modern microwave transistor may be in the tens of gigahertz range, we have a very wide frequency range where unwanted oscillation is

a concern. Accordingly, designers must exercise special care with circuit layout, dc bias networks, and enclosures.

1.5.1.1 RF instability created by unwanted feedback

To illustrate how quickly improper grounding can cause problems, let us look at an example where we wanted to extend the frequency response of two cascaded Agilent 011710 RFICs to 1 GHz by adding a matched gain equalizer between the two amplifiers. The gain of each individual amplifier rolls off from 500 MHz and it is more than 2 dB down at 1 GHz. Without equalization, the gain of two cascaded amplifiers is about 5 dB down at 1 GHz. Adding a passive equalizer between the two stages extends the frequency response to 1 GHz by giving up gain at low frequencies. Initial circuit simulation confirms the flat gain and very good input/output impedance match for the equalized two-stage, as shown in Figure 1.15. At this point, perfect groundings were used for both circuits.

When the two-stage amplifier was initially built, it was grounded by a common via hole, representing about 0.25-nH effective inductance [Figure 1.16(a)]. The effect of this seemingly small common-mode ground inductance was disastrous: a large gain peak at 1,000 MHz and input/output reflection coefficient magnitudes greater than unity. Grounding the two gain blocks separately with the same type of individual via holes, as shown in Figure 1.16(b), fixed the problem.

We cannot avoid having some minimum ground inductance, but by grounding the two stages separately, the signal current of the second stage is not fed back directly to the first stage. Even though the common-ground inductance used in our example is very small, at 1 GHz it represents $j1.5\Omega$ inductive reactance, which is enough to create harmful positive feedback at that frequency. Figure 1.17 compares the gains and input reflection coefficients of the two different grounding methods, showing the kind of problems that may be caused by improper grounding.

1.5.2 Stability analysis with arbitrary source and load terminations

Next, let us examine if there are other conditions that may lead to unwanted RF oscillation. Before going back to the two-port devices, however, let us look at an active one-port to see under what conditions oscillation may occur. Then, we will extend our discussion to the two-port case.

1.5.2.1 One-port stability considerations

In Volume I, we see that passive components always have reflection coefficient magnitudes less than unity. The real parts of their impedances are

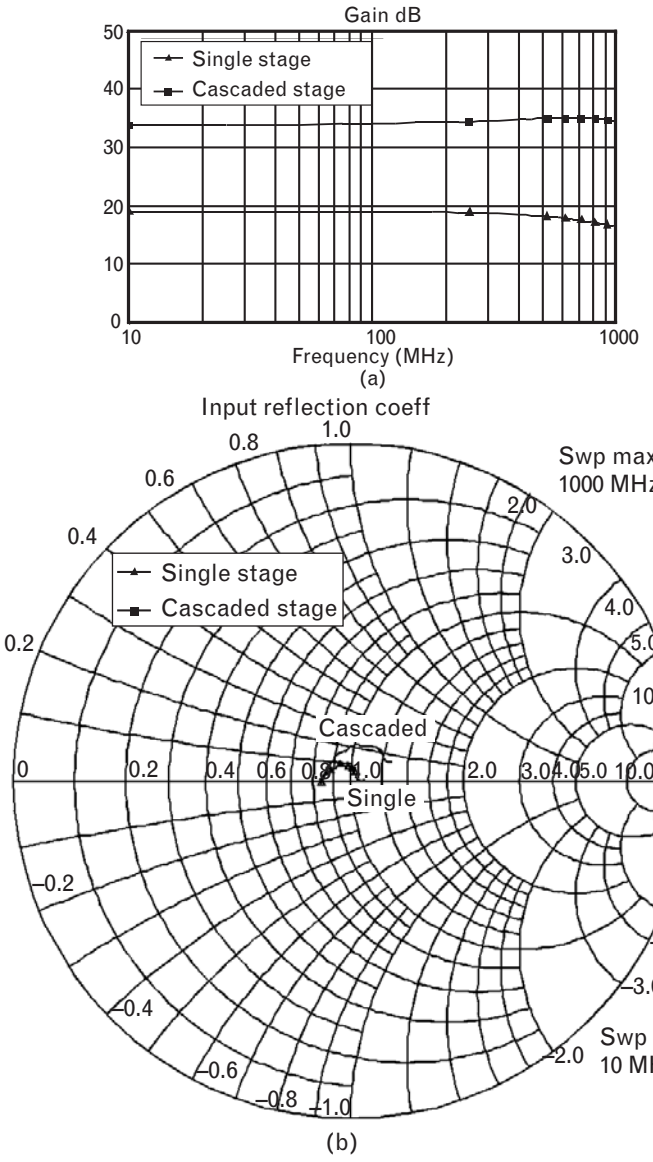


FIGURE 1.15 *Simulated (a) gain response and (b) input reflection coefficient of a single-stage and the equalized two-stage gain-module, using ideal grounding.*

FIGURE 1.16 *Circuit schematics of a two-stage amplifier module with gain equalization, showing two different nonideal RF grounding methods: (a) the two gain-blocks have a common RF grounding path represented by 0.25-nH equivalent via hole inductance, labeled as L_1; and (b) separating the grounds by using two via holes, L_2 and L_3.*

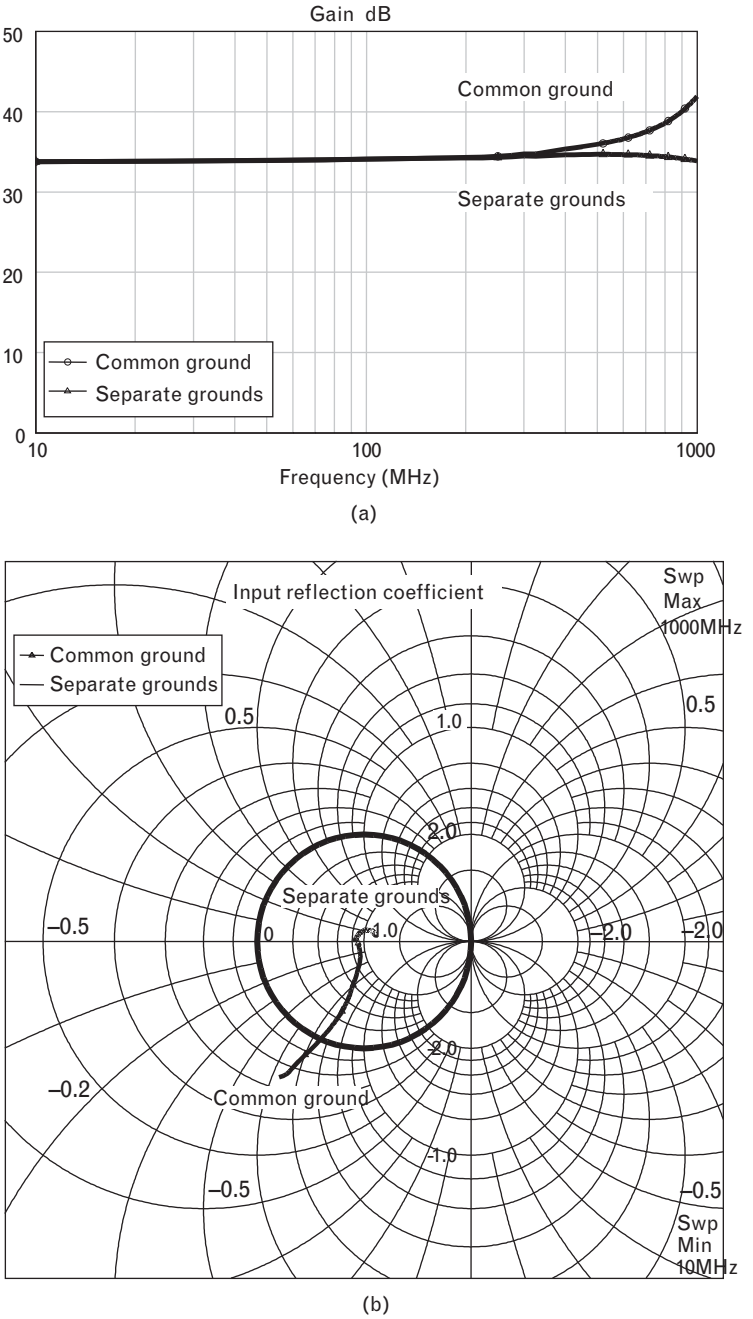


FIGURE 1.17
(a) Gain response and (b) input reflection coefficient of the equalized two-stage amplifier with nonideal common-ground inductance of 0.25 nH, and with two separate 0.25-nH inductances. The common-mode feedback, caused by the single ground path with 0.25-nH inductance, leads to a 7-dB gain peak at 1 GHz and $|s_{11}| = 1.6$. Separating the grounds [Figure 1.16(b)] reduces the maximum reflection coefficient of the cascaded amplifier to a magnitude of 0.1.

always located in the right-hand plane of the rectangular impedance system. That is a basic definition of passivity. Next, we look at active components, where reflection coefficient magnitudes may exceed unity, meaning they provide *reflection gain*. Now the reflected signal is larger than the incident signal. On the Smith chart, constant resistance circles with negative

values are always outside of the unit-radius chart, and any reflection coefficient with magnitude larger than unity refers to *negative resistance*, meaning the real part now lies in the left-side impedance plane.

Negative resistance refers to a component or a circuit where an incremental increase of the applied voltage leads to a decrease of current ($-R = \Delta v/-\Delta i$). It can only be realized by active circuit elements, such as a tunnel-diode [11] or a transistor, and it may only be negative for a specific range of the applied bias conditions and frequency range. Negative resistance can occur in a heavily doped p-n junction, having such a thin depletion layer that electrons tunnel through at relatively low forward bias voltages. The effect goes away as the dc voltage is increased. The range of negative resistance is a function of the type of semiconductors used. For example, looking at the current-voltage characteristics of the GaSb tunnel-diode in Figure 1.18, we can see the negative resistance region exists between 0.05V and 0.3V. For GaAs that range is considerably wider and occurs at higher junction voltages.

An intuition-based, though *not always correct,* explanation of how oscillation may build up between an active and a passive port is based on the reflection coefficients of the ports. In Figure 1.19, an active one-port, having an input reflection coefficient of $|\Gamma_{IN}| > 1$, is connected to a passive element with reflection coefficient $|\Gamma_S| \leq 1$. The broadband thermal noise [12] generated by the passive element travels toward the active device and is reflected with a larger magnitude, because $|\Gamma_{IN}| > 1$. When the reflected noise reaches the passive termination, part of it is rereflected toward the active device, and the process is repeated over and over again. If at a specific

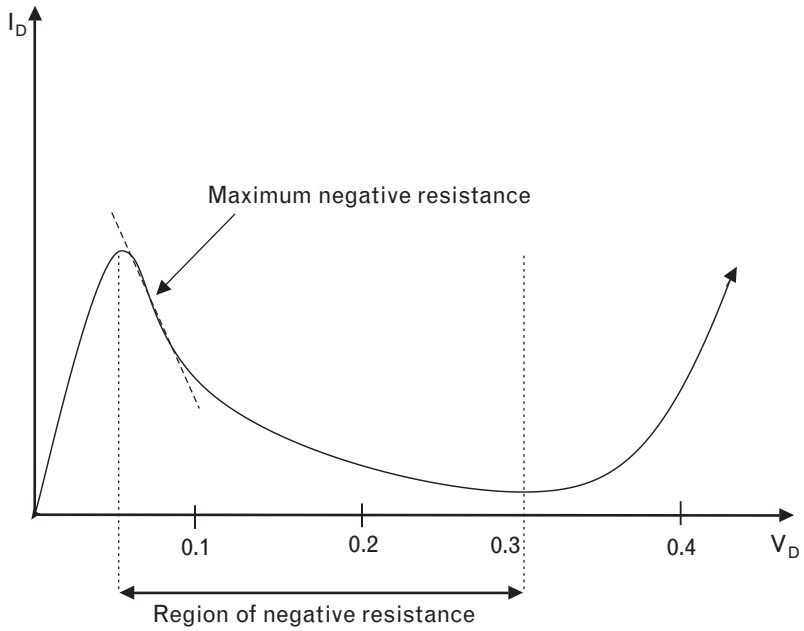


FIGURE 1.18
A tunnel-diode differs from a conventional p-n diode by having a region of negative resistance. The slope of the I-V response shows negative resistance from 50 to 300 mV.

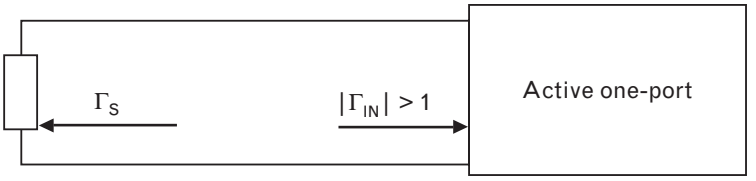


FIGURE 1.19 *The random noise generated by the termination of an active one-port may result in a buildup of oscillation if the relationship between the two reflection coefficient meets the conditions stated in (1.23) and (1.24).*

frequency the noise power increases while being bounced back and forth between the two one-ports, this may be viewed as the start-up of oscillation. Mathematically, the thermal noise voltage starts building up in the closed loop formed by the two terminations of Figure 1.19, if the loop-gain (the product of two interfacing reflection coefficients) exceeds unity, while the two phase angles cancel each other (or are multiples of 360°).

$$|\Gamma_S \Gamma_{IN}| > 1 \tag{1.23}$$

and the phase angle is

$$|\Gamma_S \Gamma_{IN}| = 0^\circ (\text{or multiples of } 360^\circ) \tag{1.24}$$

At the frequency where the above two conditions are met, the magnitude of noise rapidly increases and eventually forces the active device into its large-signal mode. At that point, Γ_{IN} begins to change, and the nature of how that change takes place determines whether steady-state oscillation will be reached or not.

This description is a highly simplified, and not always correct, explanation of how oscillation begins. As we mentioned, an accurate and more detailed discussion will follow in Chapters 4 and 6, when we will also examine large-signal considerations and the Nyquist test [13]. Only then can we really determine if steady-state oscillation will take place. In the meantime, let us stay with the small-signal analysis to see if our circuits are capable of creating the start-up conditions for oscillation.

In graphical form, the Nyquist stability criteria plots the open-loop gain function on a complex polar plane from negative infinite frequency to positive infinity. From (1.20), the loop gain function is part of the denominator. For the purpose of the Nyquist test, this loop gain function for the circuit shown in Figure 1.19 can be expressed as the product of the two reflection coefficients in the complex *s*-plane:

$$G_N = \Gamma_S(s)\Gamma_{IN}(s)$$

where $s = j\omega$.

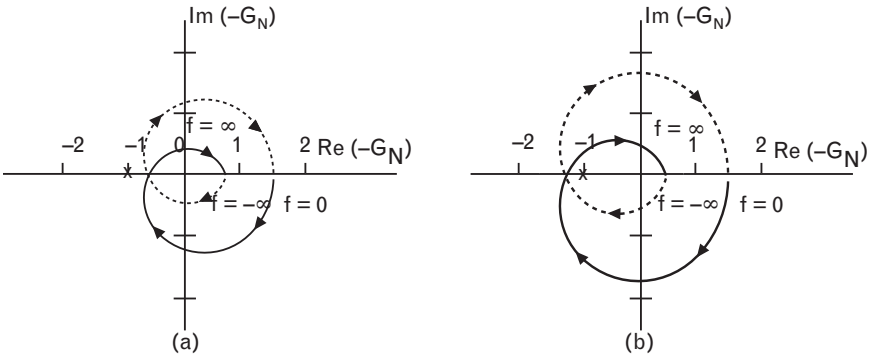


FIGURE 1.20
Examples of (a) stable and (b) unstable port connections. If the plot of $-G_N$ encircles the location of -1 of a complex plane (marked with "x") in a clockwise direction, the port connection is unstable.

The system is classified unstable if the plotted loop gain function encircles a specific point in clockwise direction. Various textbooks and CAD programs perform this test three different ways:

1. Using the loop gain function G_N to see if the point $(+1+j0)$ is encircled;
2. Using the negated value of the loop gain function, $-G_N$ as shown to see if $(-1+j0)$ is encircled;
3. Using the quantity $(1 - G_N)$ to see if the origin $(0+j0)$ is encircled.

In this chapter we use the second of the three options. If the Nyquist plot encircles –1 in a clockwise direction, the closed loop is unstable, as shown in Figure 1.20. In the oscillator discussion of Chapter 6, we will switch to the first option for comparison. The results are the same, regardless of which option is used, since testing the negated gain function for –1 location is the same as testing the positive gain function for +1. Effectively, they all test the denominator of (1.20) for poles of G_N in the right-half plane.

Stability analysis between one-ports is relatively simple. Analysis of bilateral two-port circuits is more complicated because the reflection coefficient at one side of the two-port is a function of the termination connected to the other side. In that case, the Nyquist test can only tell if the circuit is stable with the specific terminations used.

1.5.3 Two-port stability considerations

Just as in the one-port case, the two-port shown in Figure 1.21 may also start up oscillation if reflected signals, either at the input or output port, increase their magnitudes while they are continuously reflected between an active port and its termination. Such conditions often occur far below the passband frequency of an amplifier, where the transistors have high gain

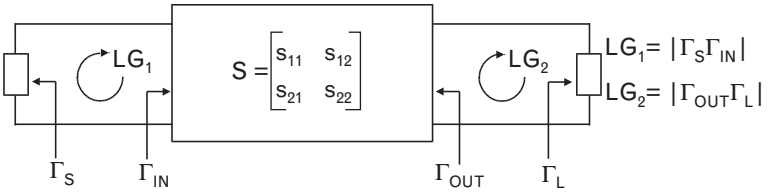


FIGURE 1.21 *An active two-port, characterized by its scattering matrix S, may oscillate if either of the loop-gain products (LG_1 or LG_2) exceeds unity. If the real parts of the source and load impedances are zero or positive ($|\Gamma_S| < 1.0$ and $|\Gamma_L| < 1.0$), then oscillatory conditions can only exist if either $|\Gamma_{IN}|$ or $|\Gamma_{OUT}|$ is greater than unity.*

and the terminations seen by the device are far from 50Ω. Antennas, filters, and couplers are good examples of such terminations.

If we restrict our terminations to those located inside the Smith chart (source and load impedances have positive real parts), in the absence of an external feedback path, oscillation can only build up if either

$$|\Gamma_{IN}| > 1$$

or

$$|\Gamma_{OUT}| > 1$$

An early form of a two-port RF stability test was defined as the ability to conjugate match a two-port simultaneously with positive real terminations, without the possibility of oscillation. Later, it was shown that simultaneous conjugate-match is not always the most conclusive test for stability. To assure stability for all possible passive source and load terminations, we must be convinced that neither Γ_{IN} or Γ_{OUT} can have magnitudes greater than unity [14].

(We should clear up here that when we use the term passive termination, it may actually represent one of the ports of an active circuit, as long as the reflection coefficient magnitude does not exceed unity. Therefore, the port-impedance of the termination has a real part with zero or positive value.)

1.5.3.1 Analytic definition of two-port RF stability—the *K*-factor

Mathematically, unconditional two-port stability exists when

$$|\Gamma_{IN}| = \left| s_{11} + \frac{s_{12}\, s_{21}\, \Gamma_L}{1 - s_{22}\, \Gamma_L} \right| < 1 \tag{1.25}$$

and

$$\left|\Gamma_{OUT}\right| = \left| s_{22} + \frac{s_{12}s_{21}\Gamma_S}{1 - s_{11}\Gamma_S} \right| < 1 \tag{1.26}$$

for all

$$|\Gamma_S| \leq 1 \tag{1.27}$$

and

$$|\Gamma_L| \leq 1 \tag{1.28}$$

From (1.25) to (1.28), we can define two requirements for two-port stability in terms of *S*-parameters. First, the stability *K*-factor,

$$K = \frac{1 - |s_{11}|^2 - |s_{22}|^2 + |\Delta|^2}{2|s_{12}s_{21}|} > 1 \tag{1.29}$$

and either one of the following two conditions [8] (there are actually are five of these secondary conditions, but we are only showing the most commonly used ones):

$$|\Delta| = |s_{11}s_{22} - s_{21}s_{12}| < 1 \tag{1.30}$$

$$B_1 = 1 + |s_{11}|^2 - |s_{22}|^2 - |\Delta|^2 > 0 \tag{1.31}$$

If the two-port satisfies both (1.29) and (1.30) or (1.31), it is classified to be unconditionally stable, otherwise it is called *potentially unstable* (sometimes referred to as *potentially stable*).

The *K*-factor of a two-port is invariant; it does not vary when lossless components are cascaded to the input or output port. However, cascaded lossy elements, lossy or lossless feedback, do change the *K*-factor.

1.5.3.2 A better stability criteria—the μ-factor

Since the stability definition takes two separate tests, it is difficult to compare the relative stability of various devices. A later development [15] combines the two tests into a single, more practical form, the μ-factor that needs to be greater than unity for stability. The μ-factor is very useful to compare the relative stability of devices: *Larger values indicate greater stability*,

$$\mu_1 = \frac{1-|s_{22}|^2}{|s_{11} - \Delta(s_{22}^{\star})|+|s_{21}s_{12}|} > 1 \tag{1.32}$$

We should explain that there are actually two μ-factors:[9] μ_1 and μ_2. Equation (1.32) is the one generally used and referred to as *μ-factor*. The second factor, μ_2, is computed from an expression similar to (1.32) by simply interchanging s_{11} and s_{22}. However, if μ_1 is greater than 1.0, then μ_2 is also greater than 1.0, so it is not necessary to compute both μ-factors.

The μ-factors also have very meaningful physical interpretations: μ_1 is the distance between the center of the Smith chart and the unstable region of the load stability circle (see Figure 1.22), while μ_2 shows how far the unstable region of the source stability circle is from the center of the Smith chart. We cover the concept of stability circles in Section 1.5.4.

The μ-factor also makes it easy to compare stability of different transistors. For example, if we have five devices with known *S*-parameters, we simply compute the μ-factor of each and rank devices in the order of their μ-factors. The transistor with highest μ-factor is the most stable one.

Viewing the signal-flow graph [16] of a two-port terminated with arbitrary impedances may help to visualize how signals travel in both directions in RF circuits (Figure 1.23). We can easily identify three loops where oscillation may start up, as follows:

1. *Input loop*, having loop-gain of ($\Gamma_S s_{11}$);
2. *Output loop*, having loop-gain of ($s_{22}\Gamma_L$);

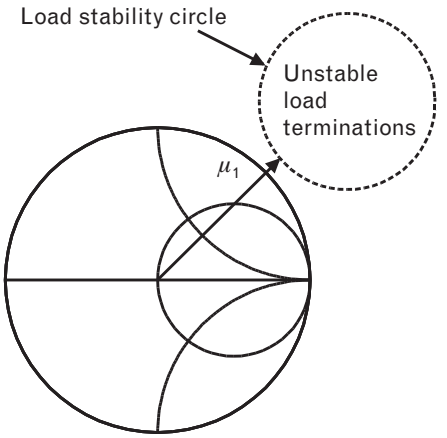


FIGURE 1.22 *While the stability K-factor is only an analytical definition, the μ-factors show exactly how far the regions of unstable terminations are from the center of the Smith chart. If the magnitudes of the μ-factors are greater than unity, then any termination on the Smith chart may be used safely. This illustration shows the definition of μ_1, generally referred to just as μ,*

9. In the original publication the authors used μ and μ'.

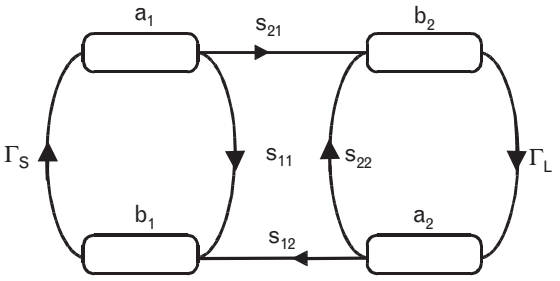


FIGURE 1.23 *Signal flow-graph illustrates the direction of forward and reflected waves through and around a two-port terminated with an arbitrary source and load. Nodes a_1 and b_1 refer to the input port, while a_2 and b_2 refer to the output port. Oscillation may start up within one of the three loops of the graph if certain conditions are met.*

3. *Overall feedback loop*, having loop-gain of ($\Gamma_S\, s_{21} \Gamma_L\, s_{12}$).

Excessive magnitude of a loop gain with the wrong phase shift may lead to oscillation. Limiting the magnitudes of Γ_S and Γ_L to a maximum of 1.0, the loop gain of the largest loop is strongly influenced by the $s_{21}s_{12}$ product. Later, while looking at RF stabilization methods, we will see that resistive attenuation of this loop improves stability.

Since all stability tests are based on frequency-dependent small-signal *S*-parameters, it is easy to see that two-port stability changes with frequency. Generally, active devices are stable at the very low frequencies where $|s_{12}|$ is very small, and also at the very high frequencies where $|s_{21}|$ rolls off. Unfortunately (for amplifier designers) there is a wide range of RF/MW frequencies where the possibility of oscillation is a threat to stable operation, as indicated in Figure 1.24.

The stability factor is also a function of dc bias settings and the signal level. When the applied signal level begins to compress the gain of the device, the *S*-parameters change and so does the stability factor.

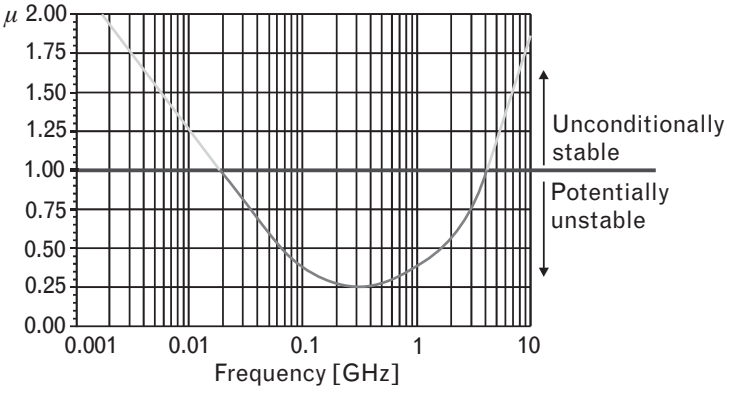


FIGURE 1.24 *Broadband stability characterization of a typical RF transistor, showing potential instability in the 20- to 4,000-MHz frequency range where $\mu < 1$.*

Let us point out again that although the *K*-factor is not affected by any lossless component cascaded to the two-port, the μ-factors do change for the same. However, if either one has less than unity value, only feedback or lossy cascaded elements can increase the magnitude above unity.

1.5.4 Stability circles

The *K*-factor and μ-factor help us to classify a two-port as stable or potentially unstable. To stabilize the two-port, we need to know what type of terminations can lead to possible oscillation.

If a two-port is potentially unstable, then:

- There are *unfriendly source terminations* for which the magnitude of the *output reflection coefficient* becomes greater than unity.
- There are also *unfriendly load terminations* for which the magnitude of the *input reflection coefficient* becomes greater than unity.

For our discussion here we will temporarily introduce the terms *friendly* and *unfriendly* terminations. An unfriendly termination is the type that leads to reflection coefficients with greater than unity magnitude and possibly to unwanted oscillation. Friendly terminations, on the other hand, keep the reflection coefficient magnitudes under unity, preventing oscillation. Between the friendly and unfriendly terminations, we define a third category of *borderline terminations* that result in exact unity reflection coefficient magnitudes at the adjacent port. The three newly defined terminations are illustrated for the source side in Figure 1.25. The same three possibilities apply for the load terminations also.

Keep in mind, however, that oscillator designers will disagree with these definitions, because they want to create oscillation. They want reflection coefficients to *exceed* unity magnitudes. In Chapter 6 we will discuss how to take advantage of existing potential instability, and even how to create instability for oscillator design.

Virtually all commercially available RF/MW transistors exhibit potential instability at some frequencies. In order to stabilize a device, we first need to identify the unfriendly terminations that may lead to oscillation, and then stabilize the device by adding a protective circuitry, so that it cannot directly interface any unfriendly termination

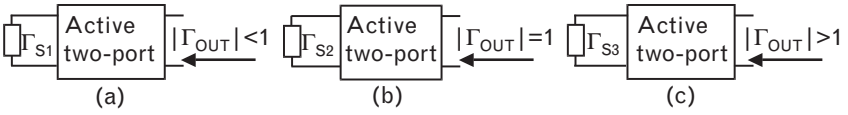


FIGURE 1.25 *Three types of source terminations leading to different reflection coefficient magnitudes at the output port: (a) Γ_{S1} is friendly because it leads to $|\Gamma_{OUT}| < 1.0$; (b) Γ_{S2} is borderline because $|\Gamma_{OUT}| = 1.0$; and (c) Γ_{S3} is unfriendly, since it causes $|\Gamma_{OUT}| > 1.0$.*

Since a picture is worth a thousand words, we again turn to a Smith chart–based graphical technique for help. A visual illustration of RF stability is done through the stability circles [7], where the circumference of the circles represents the locus of all borderline terminations. Accordingly, a stability circle is the border between all stable and unstable terminations. At each frequency we can find two stability circles, one for the source terminations and another for the loads. Let us see how we find such a circle on the source side first.

Since we have an expression that relates Γ_S to Γ_{OUT}, we can now find out what kind of Γ_S leads to unity magnitude for Γ_{OUT}. We therefore set the output reflection coefficient expression (1.5) to unity and solve it for Γ_S.

For convenience, we first rewrite (1.5),

$$\Gamma_{OUT} = s_{22} + \frac{s_{12}s_{21}\Gamma_S}{1 - s_{11}\Gamma_S}$$

and set $|\Gamma_{OUT}| = 1.0$

$$\left| s_{22} + \frac{s_{12}s_{21}\Gamma_S}{1 - s_{11}\Gamma_S} \right| = 1 \tag{1.33}$$

Solving (1.33) for Γ_S provides the equations of a circle, called the *source stability circle*, shown in Figure 1.26. The center of the circle is located at

$$C_S = \frac{(s_{11} - s_{22}^{*}\Delta)^{*}}{|s_{11}|^2 - |\Delta|^2}$$

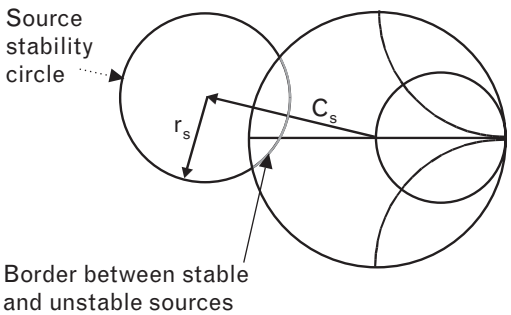


FIGURE 1.26 *Source stability circle of a two-port plotted on the Smith chart. C_S is a complex number, representing the center, and r_S is the magnitude of the circle. Connecting any source located on the circumference of the stability circle to the input of the two-port results in* $|\Gamma_{OUT}| =$

and the radius is

$$r_S = \frac{|s_{21}s_{12}|}{|s_{11}|^2 - |\Delta|^2}$$

Similarly, from (1.3), setting the magnitude of Γ_{IN} equal to unity gives

$$\left| s_{11} + \frac{s_{12}s_{21}\Gamma_L}{1 - s_{22}\Gamma_L} \right| = 1$$

The values of Γ_L that satisfy this equation give us the load stability circle. A typical plot of both stability circles is shown in Figure 1.27. The centers (C_S and C_L) and radii (r_S and r_L) of the circles are computed from the two-port *S*-parameters, and the circles can be plotted by most of the RF/MW CAE programs.

Interpretation of the stability circles would be quite straightforward if they would consistently indicate the stable and unstable regions. Unfortunately, there are cases when the inside region of a circle refers to the stable terminations, and other times to unstable terminations. Conditions can also change from one frequency to another. However, there is always a simple, intuitive way to select the proper region, as outlined in the following section. Even though our explanation refers only to the source (input) stability circles plotted on a 50-Ω normalized Smith chart, the same reasoning can also be applied to the load circles, at the output side of the two-port.

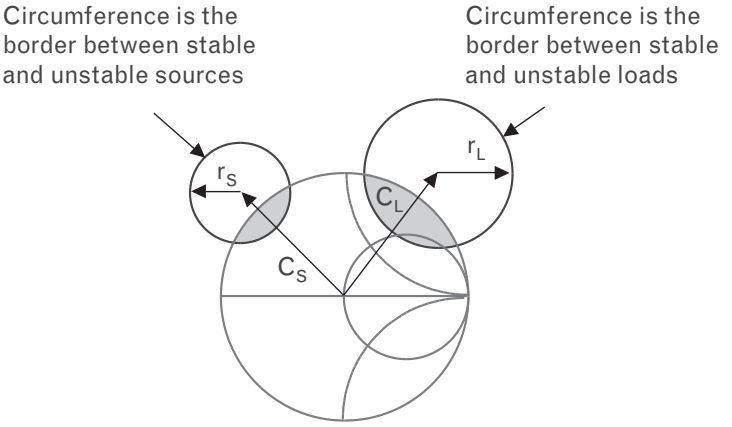


FIGURE 1.27 *Single-frequency source and load stability circles of a typical RF transistor, indicating potential instability. Since the circles intersect the Smith chart, a portion of the chart contains*
terminations that could lead to oscillation. If the inside of the stability circles indicate the unstable

1.5.4.1 How to determine the stable side of a stability circle

The circumference of a source stability circle is the locus of all source terminations that forms a border for the stability considerations of the output port. This is a very important and often not a clearly understood point, so let us restate the three fundamental conditions again. When $|\Gamma_{OUT}| > 1.0$, oscillation may take place at the output port. If $|\Gamma_{OUT}| < 1.0$, we have a stable output port. Between these two extremes lies the borderline case of $|\Gamma_{OUT}| = 1.0$.

The next step is to determine if the region inside the stability circle represents stable (friendly) or unstable (unfriendly) terminations. Although the common belief is that the inside region is the stable one, it is not always true. Most of the commercial RF/MW circuit simulators indicate or label the stable region. There is also a simple test for this question, which is described next.

We need to select a source termination that is *not* on the circumference of the Smith chart and investigate whether it causes $|\Gamma_{OUT}|$ to be less than unity (stable) or greater than unity (potentially unstable). An obvious choice for this test is 50Ω, the original source used during the initial *S*-parameter measurements, since that was used while the basic s_{22} of the device was measured. To rephrase, we want to know if the magnitude of the output reflection coefficient is larger than unity when the source is equal to 50Ω. If $|s_{22}| < 1.0$, then a 50-Ω source is *friendly*, or the *stable* type. When $|s_{22}| > 1.0$, then 50Ω is classified as *unfriendly* and may lead to oscillation. Figure 1.28 illustrates the two possible ways to determine which side of the source stability circle is stable. For simplicity, we use the same stability circle plots for both cases.

Since we are making a decision about the source terminations, we need to ask: What is the magnitude of the output reflection coefficient? Besides the borderline case ($|\Gamma_{OUT}| = 1.0$), there are two possibilities: The magnitude is less than unity or greater than unity.

- *Case A:* If $|s_{22}| < 1.0$, then the 50-Ω source is classified as a termination leading to *stable* output.
- *Case B:* When $|s_{22}| > 1.0$, then the 50-Ω source leads to a potentially *unstable* output.

1.5.4.2 Illustrative example: finding the unstable source termination region

Given: The input reflection coefficients of four different transistors with their corresponding source stability circle plots. Identify the unstable source regions for all.

Solution: Other than the borderline case, there are four possible combinations, since the source stability circle may or may not enclose the center

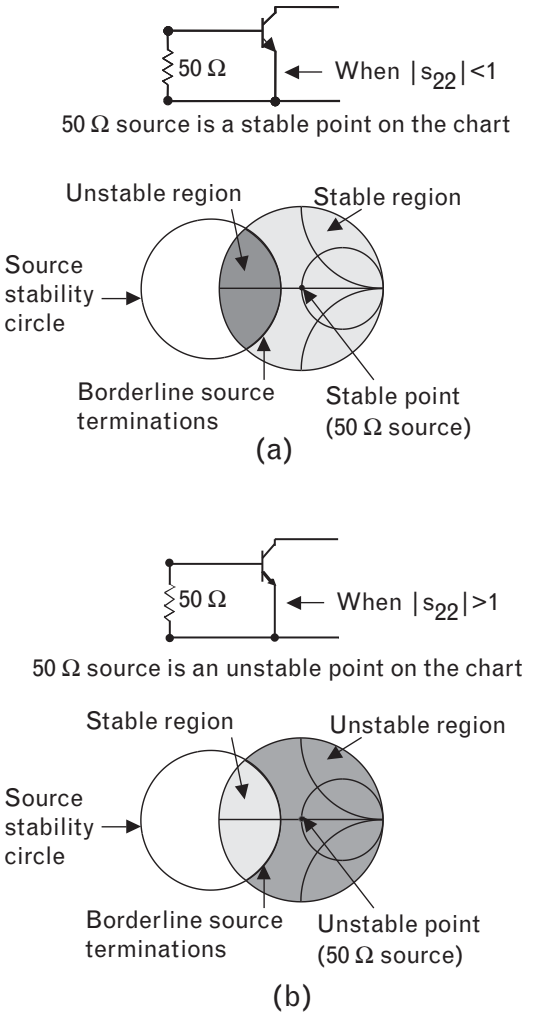


FIGURE 1.28
If the device S-parameters were measured between two 50-Ω terminations, $|s_{22}|$ indicates how the 50-Ω source impedance, located at the center of the Smith chart, should be classified. (a) If $|s_{22}| < 1$, the center of the Smith chart is stable; and (b) if $|s_{22}| > 1$, the center is unstable.

of the Smith chart, and $|s_{22}|$ may be greater or less than unity. Therefore, using the procedure outlined in Section 1.5.4.1, we can label the 50-Ω point (center of the Smith chart) accordingly stable or potentially unstable, and see if that point is inside or outside of the stability circle. As soon as this point is labeled, both sides of the stability circle can also be labeled either stable or unstable, as illustrated in Figure 1.29. (*Note:* If the circumference of a stability circle crosses the center of the Smith chart, we have an ambiguous case, and we let the mathematicians decide the outcome.)

Solutions

We first look at $|s_{22}|$ to find out if a 50-Ω source is a stable or unstable point on the Smith chart.

(a) $s_{22} = 0.50 \angle -140°$

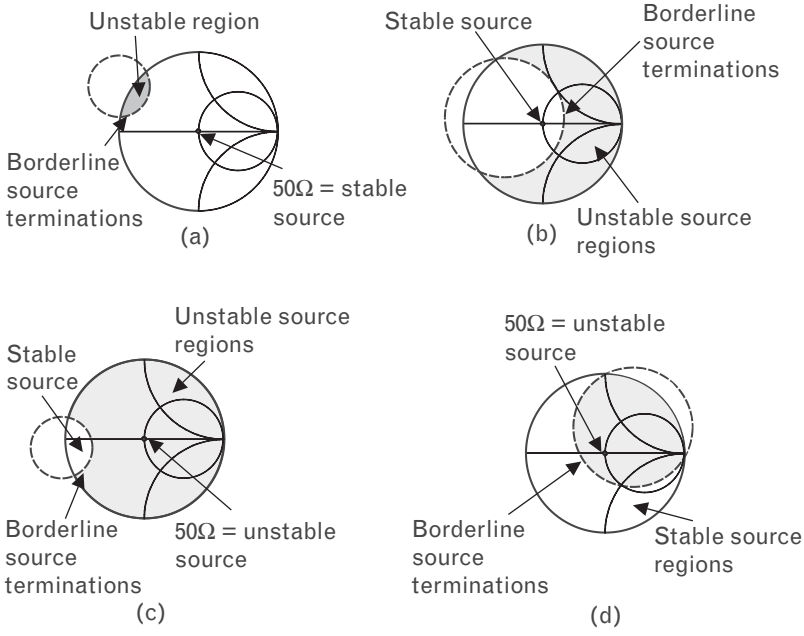


FIGURE 1.29
The original $|s_{22}|$ of the device may either be less or greater than unity, resulting in four possible classifications. Unstable sources are indicated by the shaded regions. Parts (a) through (d) refer to the four cases outlined under solutions.

$|s_{22}| < 1.0$, and the source stability circle does not enclose the center of the Smith chart (center is a stable source). *Conclusion:* inside region unstable, outside stable.

(b) $s_{22} = 0.66 \angle -23°$

$|s_{22}| < 1.0$ and the source stability circle encloses the center of the Smith chart (center is a stable source). *Conclusion:* inside region stable, outside unstable.

(c) $s_{22} = 1.12 \angle 170°$

$|s_{22}| > 1.0$ and the source stability circle does not enclose the center of the Smith chart (center is an unstable source). *Conclusion:* inside region stable, outside unstable.

(d) $s_{22} = 1.20 \angle -35°$

$|s_{22}| > 1.0$ and the source stability circle encloses the center of the Smith chart (center is an unstable source). *Conclusion:* inside region unstable, outside stable.

1.5.5 Graphical forms of unconditional stability

In the preceding example, we saw that the inside of the source stability circle indicates either the stable or the unstable source terminations. For unconditional stability, all terminations within the Smith chart must be declared stable. There are two possible forms, as shown in Figure 1.30.

- In Figure 1.30(a) the stability circle is outside of the Smith chart and the inside of the circle is the region of unstable terminations.

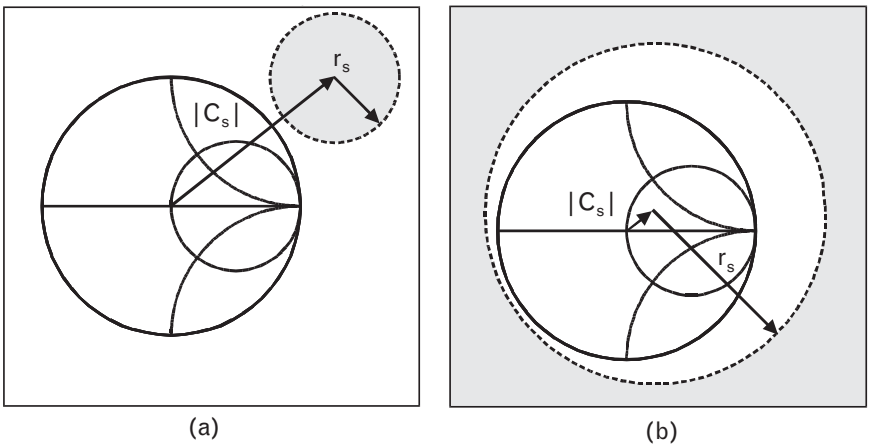


FIGURE 1.30
(a, b) Two forms of stability for source circles. The shaded areas show the region of unstable sources. In both illustrations the Smith chart is in the stable region. Stability circles are defined by two parameters: the radius, r_s, and the center vector, C_s, computed from the S-parameters of the two-port.

- In Figure 1.30(b) the stability circle completely encloses the chart and the inside of the stability circle shows the stable region.

In both cases the complete Smith chart represents the stable region, so all passive terminations may be chosen freely without the risk of oscillation. Once again, we only show the source circle here and for complete stability analysis we must check both the source and load circles. If both ports are stable, then the device is classified as *unconditionally stable*.

1.5.6 Graphical forms of potential instability

A stability circle may indicate potential instability in three ways, as shown for the source circle in Figure 1.31. The stability circle may:

1. Intersect the Smith chart;
2. Be placed entirely inside the Smith chart;
3. Be outside the Smith chart in one of the following forms:
   Enclose the chart and the inside of the circle represents the unstable region;
   Be away from the chart and the outside of the circle represents the unstable region.

If the unstable region of terminations includes *any portion* of the Smith chart, the two-port is potentially unstable. However, potential instability does not mean that the two-port oscillates—it only indicates the ability to generate negative resistance at one of the ports, which later may lead to oscillation when the improper termination is used.

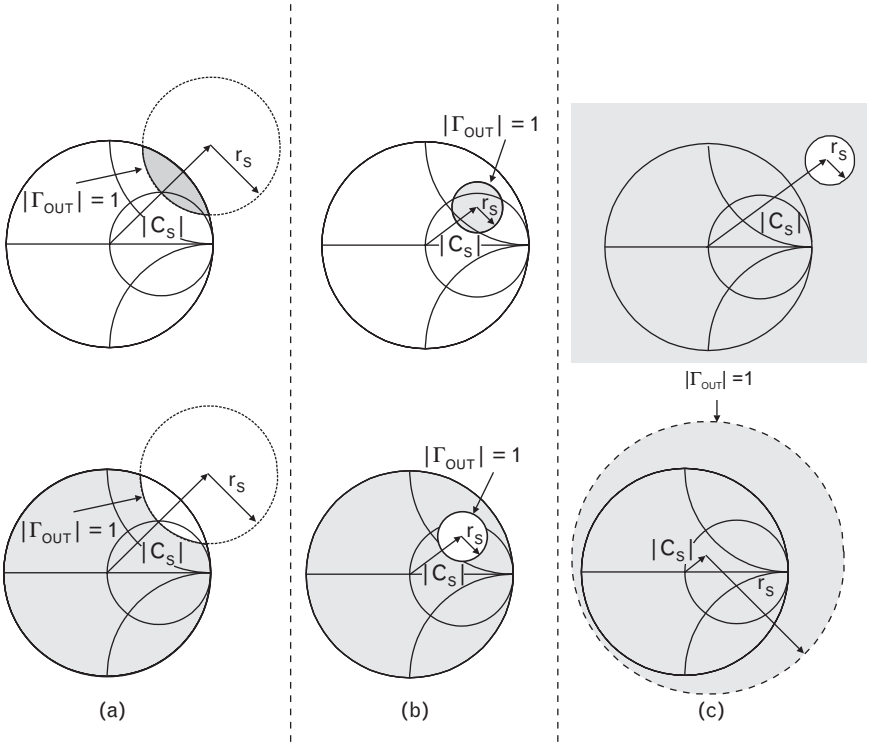


FIGURE 1.31
Three possible forms of potentially unstable source terminations: (a) intersecting, (b) inside, and (c) outside the Smith chart. Clearly, in all three sets of plots, the shaded parts of the Smith charts refer to terminations that may lead to oscillation; (c) is the worst case).

1.5.7 Caution about multistage systems

In the case of multistage amplifiers, the overall two-port stability factor, computed from the *S*-parameters of the two-port, tests:

- The input reflection coefficient magnitude of the first stage as a function of the load connected to the last stage;
- The output reflection coefficient magnitude of the last stage as a function of the source connected to the first stage.

When every stage of the cascade is unconditionally stable, the overall circuit is also stable unless some form of feedback (i.e., RF leakage through poorly filtered dc bias circuitry or radiation) creates oscillatory conditions.

If one or more stages of a cascade are potentially unstable, or if improper feedback exists, oscillation may start up if the complete circuit's μ-factor is less than unity. The overall circuit's stability circles can warn us about the regions of troublesome terminations and the frequencies. When a potentially unstable multistage circuit sees the wrong termination(s) at a specific frequency, oscillation may start at the input or output port, or in one of the interstages, depending on additional conditions covered in Chapter 6.

While *S*-parameter analysis of the cascade is straightforward, it is not easy to include all the feedback and coupling (inductive or capacitive) effects because they are often not obvious and may require additional design methodology, such as *electromagnetic* (EM) simulation. Analyzing the interstages of multistage circuits [17] may also be difficult when a common dc bias source feeds every stage.

Table 1.3 shows the stability analysis of two cascaded transistors that are potentially unstable at several frequencies. Looking at the overall stability factor of the two-stage, we conclude that the circuit is unconditionally stable, since the μ-factor of the overall two-port is greater than unity at all listed frequencies. However, adding a short segment of 50-Ω transmission line between the two stages completely changes the picture. The two-port is now potentially unstable between 0.5 and 3 GHz, but much more stable at 6 GHz, showing how sensitive the components are to what is used for the interconnection.

As we see by looking at the μ-factors listed in Table 1.3, cascading two potentially unstable two-ports can lead to an unconditionally stable or a potentially unstable circuit, depending on the form of interconnection. In the two-stage cascade, the output impedance of the first stage represents the source to the second stage. Similarly, the input of the second stage is the load to the first stage. When the two stages are directly cascaded to each other, neither stage represents termination to the other to create potential instability at the input or the output port of the cascade.

Adding the transmission line between the devices transforms the phase angles of the interstage impedances. Now, at some frequencies, the

Table 1.3 Tabulated μ-Factors of Two Bipolar Transistors Individually (BFP 405 at 2V, 2 mA and BFP 640 at 2V, 20 mA) and in Cascaded Forms

Frequency (GHz)	BFP 405 Alone	BFP 640 Alone	Two-stage with Direct Connection	Two-stage with Transmission Line
0.05	0.923	0.827	1.030	1.076
0.1	0.852	0.691	1.316	1.254
0.5	0.566	0.513	2.034	0.791
1	0.508	0.705	1.991	0.966
2	0.636	0.932	1.799	0.768
3	0.787	1.022	1.695	0.794
4	0.905	1.065	1.570	1.268
5	0.999	1.074	1.444	2.681
6	1.096	1.073	1.374	6.303

Note: The single-stage μ-factors are less than unity at most frequencies. Yet the overall circuit's μ-factor is always greater than unity when the devices are connected directly. A short transmission line segment between the two transistors creates significant changes in stability.

terminations are in the unstable regions. As a result, between 0.5 and 3 GHz the cascaded circuit shows potential instability.

In the absence of single-stage stability information, a multistage circuit may still be simulated for stability by adding the so-called *S*-probes [18] into each interstage and by following the special analytical test outlined in the cited reference. The *S*-probe technique—a simplified form of loop-gain analysis—allows us to investigate multistage feedback effects caused by component coupling and dc bias circuitry without breaking the interstage loops (see Section 1.5.7.1). It is also useful to investigate stability when active terminations are used.

1.5.7.1 Illustrative exercise: interstage stability analysis of cascaded transistors

The two cascaded transistors listed in Table 1.3 show potential instability at 500 MHz ($\mu = 0.791$) when cascaded with an added transmission line segment. Test the interstage stability of the cascade for the Nyquist criteria with $\Gamma_S = \Gamma_L = 0$. Repeat the test, using $\Gamma_S = \Gamma_L = 0.99 \angle 25°$, representing high-Q inductive terminations. Use $Z_0 = 50\Omega$.

Solution: The cascaded two-ports with a short segment of 50-Ω transmission line between them are shown in Figure 1.32. Initially, source and load terminations are 50Ω. Placing a GPROBE[10] element of the MW Office program into the interstage allows us to check stability. Terminals 1 and 2 of the GPROBE element are connected to the output port of Stage 1 and input port of Stage 2. Terminals 3 and 4 provide current and voltage information to perform the Nyquist test. Although our simulation is only done from 0 to 2,000 MHz, it provides sufficient information about stability. *Note:* The CAD program does not plot the response for negative

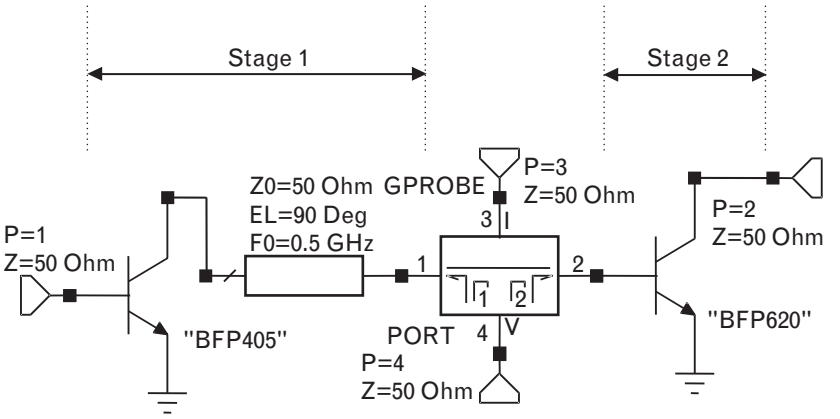


FIGURE 1.32
Setup of the Nyquist stability test, using the MW Office program. (Other RF simulators also have similar capabilities.) The 50-Ω transmission line between the two two-ports was arbitrarily chosen to represent the interconnecting link.

10. GPROBE uses the principles of the *S*-probe concept mentioned earlier.

frequencies, so we need to create a visual image of mirror-symmetry about the frequency (x-) axis.

Displaying the Nyquist plot indicates stability throughout the frequency range since the –1 location of the complex polar plane is not encircled [Figure 1.33(a)]. The test assures us that the circuit will not oscillate under this set of operating conditions—with 50-Ω terminations. However, changing the resistive terminations to inductive type ($\Gamma_S = \Gamma_L = 0.99 \angle 25°$, with $Z_0 = 50\Omega$) creates completely different results [Figure 1.33(b)]. The –1 location is now completely encircled in the clockwise direction, which is a warning sign of possible oscillation.

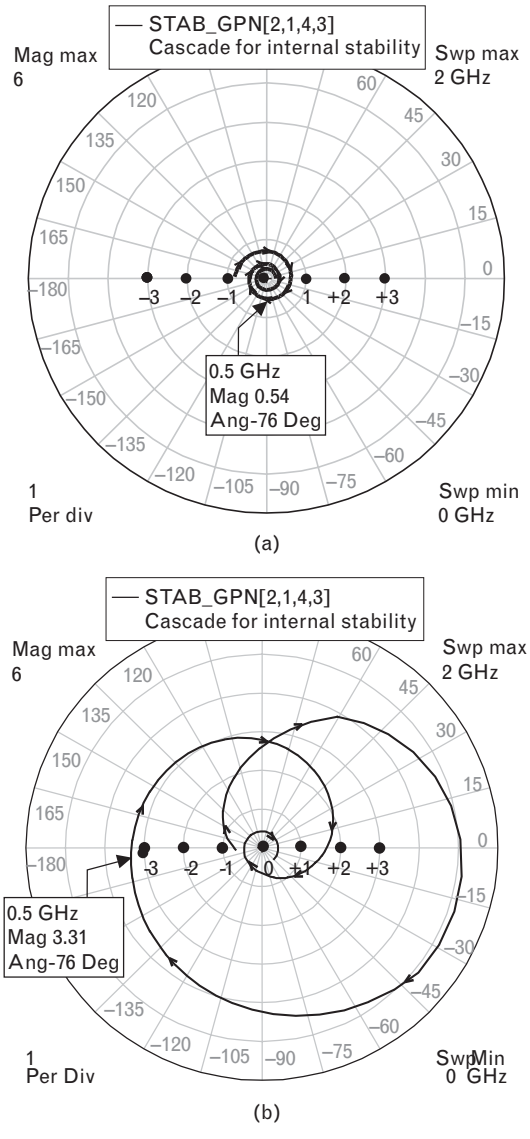


FIGURE 1.33
Two different Nyquist plots derived for the same circuit, when analyzed with two different sets of source and load terminations: (a) stable operation with $\Gamma_S = \Gamma_L = 0$; and (b) unstable results with $\Gamma_S = \Gamma_L = 0.99 \angle 25°$. Important conclusion: The Nyquist test is termination dependent. Note: Plots are shown for frequencies of 0–2,000 MHz in a clockwise direction.

Again, from small-signal analysis, we cannot tell if steady-state oscillation will take place. (Chapter 6 looks into oscillator analysis with large-signal parameters.) Still, a small-signal steady-state *S*-parameter–based test, like the one we used here, can be very helpful to identify *potential* stability problems.

The lesson to learn here is that even though in this example we have a potentially unstable two-port, it will not oscillate when the terminations are 50Ω. If the terminations at 500 MHz behave inductively, oscillation may take place, depending on the large-signal behavior of the transistors. By the way, even the Nyquist test only provides an answer for a specific set of terminations.

Is it possible that an amplifier is terminated inductively? The answer is yes. For example, if the passband of the amplifier is 2.2 to 2.3 GHz, and it is placed between two filters, most likely the filters act reactively out of their passband. It is then possible that at a frequency where the two-port is potentially unstable, the filters may represent terminations that lead to oscillation.

1.6 Stabilizing an active two-port

During my undergraduate engineering studies, I, Les Besser, had a summer job in the R&D laboratory of a large company's oscilloscope division. My task was to design the differential input circuitry of a new 100-MHz bandwidth oscilloscope. The previous two generations had 40 MHz and 10 MHz bandwidths, respectively.

Not having much experience with circuit design, I wanted to follow the circuitry of the 40-MHz scope since it was a highly successful product. Looking at the circuit schematic, I noticed that there was a 150-Ω resistor in the input terminal of each channel, for no obvious reason. I asked my project leader what function the resistors served, and he said, "Well, they have something to do with oscillation. Ask the man who designed it." When I tracked down the design engineer and asked him the same question, his reply was, "I don't exactly know why, but without the resistors the scope would oscillate occasionally, particularly with low-impedance terminations at the input." I asked him how he choose the 150-Ω value, and he said, "That's what we used in the 10-MHz scope, and it worked just fine." Not wanting to break precedent, I also added 150-Ω resistors to the inputs of my project and as far as I know, that is how the scope was produced.

While it is true that adding resistors usually helps stability, there is no guarantee that a series resistor always works. Of course, there is nothing magical about the 150-Ω value either.

This true story took place a long time ago, before the CAD era started. In my experience, however, too many superstitions and unexplained fixes still exist in the industry. They are often passed on from one to another in design laboratories and in production departments.

Potentially unstable devices can always be stabilized by cascading an appropriate *series* or *parallel* resistor—an approach that is simple and effective. Adding a dissipative element throws away transducer gain and, depending on whether the resistor is applied to the input or output, also sacrifices noise figure or output power. Obviously, we want to add enough resistance to stabilize the circuit without giving away much of the desired performance. The minimum-loss resistance refers to the resistor value that leads to a borderline stability where μ reaches unity value. Generally speaking, a higher level of stability can be achieved by adding more loss to the device.

Depending on the input-output phase relationship of the device, resistive feedback could also help, and it may be preferable over the brute force of cascading a resistive component. Application of lossless feedback may also improve stability and at the same time control other parameters, such as the optimum noise source reflection coefficient and conjugate input match [19, 20].

1.6.1 Finding the minimum-loss resistor at the input of the device

The minimum-loss cascade stabilizing resistor value can easily be determined from the Smith chart by finding the constant resistance or constant conductance circle that is tangent to the appropriate stability circle. To illustrate the process, we show the two possible choices of stabilization for the device whose source stability circles are previously shown in Figure 1.29(a). For this device, the inside region of the source stability circle indicates the unstable region. The constant-resistance and constant conductance circles that are tangent to the stable side of the stability circle indicate the minimum-loss normalized series resistance and parallel conductance, indicated in Figure 1.34.

Adding a series resistor with a normalized value of r_{SMIN} [Figure 1.34(a)] to the input guarantees that the device cannot see any source termination with resistance less than r_{SMIN}. Since all source terminations leading to instability have less than r_{SMIN} real parts, the two-port is now stabilized at the frequency where the stability circle was computed.

An alternative approach is to use a parallel conductance of g_{PMIN} value, as shown in Figure 1.34(b). Now we protect the device from seeing sources with normalized conductance less then g_{PMIN}, which again eliminates the possibility of an unstable source being connected to the input.

The effects of series or parallel stabilization are exactly the same on gain and stability. After adding either r_{SMIN} in series or g_{PMIN} in parallel, the stability

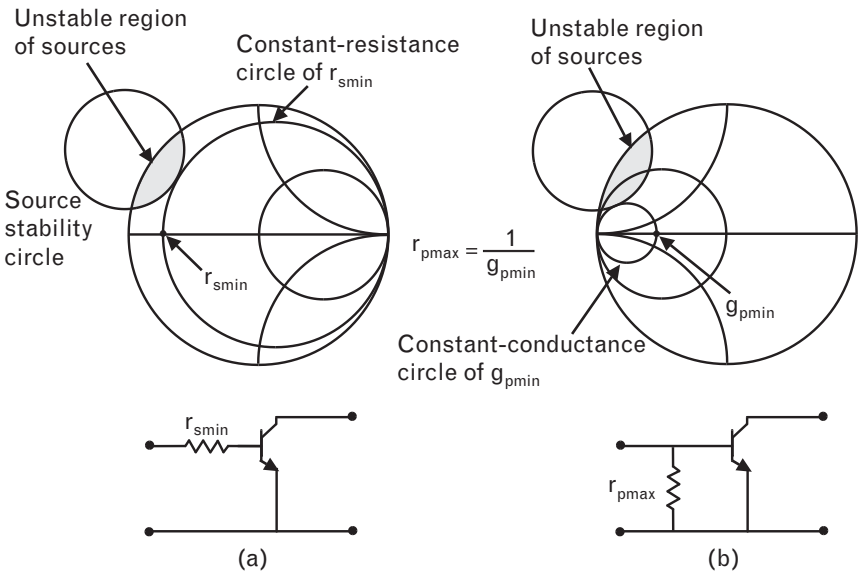


FIGURE 1.34 *Using the source stability circle to stabilize a potentially unstable device at the input port (a) with a series resistor $r_S > r_{SMIN}$, or (b) with a parallel conductor $g_P < g_{PMIN}$ (equivalent to $r_P < r_{PMAX}$). The tangent constant resistance or constant conductance circle determines the minimum loss: r_{SMIN} or g_{PMIN}. Increasing r_S or decreasing r_P leads to a greater stability margin by giving up more gain. Similar selections can be made at the output port by using the load stability circle. All resistances and conductances are normalized values.*

factor increases to unity and the effective maximum gain is the same for both forms. In practice, we generally increase the minimum value of the series resistor (decrease the parallel resistor) by 10% to 20% for an added margin of stability.

In Figure 1.34, stabilization was applied at the input of the device, and, depending on the signal and noise level of the amplifier, it may be better to stabilize at the output. Sometimes splitting the loss between the input and output leads to the best system performance. Adding the appropriate amount of minimum loss to the input or output *stabilizes both sides* of the device.

Occasionally, depending on the locations of the stability circles, series or parallel resistive stabilization is not available. In such cases, the corresponding tangent constant resistance or constant conductance circle cannot be drawn at the stable side of the stability circle. For example, in Figure 1.29(d) we saw a case where only parallel resistance (conductance) helped because open circuit (infinite impedance) is among the unstable terminations. Consequently, no matter what value of series stabilizing resistor is added, if the input port is left unterminated (i.e., open circuited), the output reflection coefficient's magnitude exceeds unity.

As a general rule, we can state that when the open circuit point of the Smith chart is unstable then series stabilization is not available; for unstable short circuit points, parallel resistance does not help.

1.6.2 Broadband stability considerations

We mentioned earlier that stabilizing the device for all frequencies, even outside the passband of interest, is a good and safe practice. Adding a stabilizing resistor degrades the performance at all frequencies, and in many cases a frequency-selective stabilizing network may be the better choice. Simple R-L or R-C combinations may sacrifice performance only where it is necessary to improve stability, without affecting other frequencies where the device may already be stable.

Figure 1.35 shows three examples of the large number of available multielement stabilization networks. The parallel R-C circuit in Figure 1.35(a) reduces the excessive gain of the active device at lower frequencies, thereby improving stability. In Figure 1.35(b), the parallel resonant circuit opens the branch at a desired frequency and lets the resistor cut the gain at both low and high frequencies. Finally, in Figure 1.35(c), the short-circuited parallel stub represents an open circuit at its quarter-wave frequency, f_R. Therefore, at that frequency the added resistor has no effect, and the branch does not create any loss. At the lower frequencies the effect of the stub becomes less significant and the resistor helps to dissipate the unwanted gain. The total impedance of this branch cycles between values of R and an open circuit through every quarter-wavelength frequency: $Z = R$ at dc, $Z = \infty$ at f_R, $Z = R$ at $2f_R$, and so on.

A thorough stability analysis should always be the first step performed before designing active circuits. Realistically, broadband unconditional amplifier stability, though desired, may not always be practical due to the physical behavior of the stabilizing elements. Still, knowing the region of terminations that may possibly lead to oscillation could provide valuable information to the designer.

Let us summarize our stability considerations:

1. Always perform a thorough stability analysis. If the source and load terminations are known for a wide range of frequencies, perhaps you can justify the analysis with those terminations only. In most

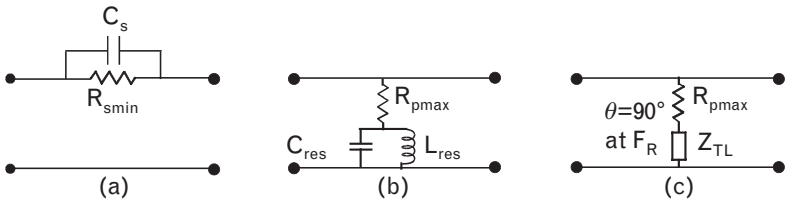


FIGURE 1.35 *Resistive stabilization dissipates RF power at all frequencies. Using complex impedance networks helps to reduce the loss at frequencies where the device stability is better. Circuits (a) and (b) are made with lumped elements, while circuit (c) uses a combination of lumped and distributed components. Resistors R_{SMIN} and R_{PMAX} represent the minimum loss required for*

cases, however, those terminations are not known, particularly outside of the passband of the amplifier.

2. Remember that the *S*-parameter–based stability analysis is not just frequency dependent. The dc bias, temperature, and high signal levels also affect stability.
3. In the case of multistage amplifiers, perform the stability analysis on the individual stages before cascading them. Overall stability analysis may provide misleading information.
4. If possible, stabilize the active devices by cascading appropriate lossy networks, using feedback, or a combination of the two. Choose the option most suitable for broadband performance.
5. If a single-stage two-port's μ-factor exceeds unity at all frequencies, the circuit cannot oscillate, as long as there is no external coupling to provide undesirable feedback. On the other hand, a smaller-than-unity μ-factor alone does not guarantee oscillation. How the terminations are chosen and what the large-signal behavior of the active device is determine whether steady-state oscillation is reached.
6. For a specific set of terminations, a Nyquist test using a true nonlinear model for the active device can tell us if the circuit will oscillate or not. Unfortunately, even the Nyquist test is termination and bias dependent and it may be difficult to model the actual terminations for a broad range of frequencies.

1.7 Stabilization of a bipolar transistor

Let us now put into practice what we have discussed by stabilizing a small-signal bipolar transistor for all the frequencies where data is available. We selected a new generation SiGe device BFP 640, from the Infineon *S*-parameter data bank at 2-V, 20-mA dc bias condition. The manufacturer provides measured data between 0.1 and 20 GHz. In Chapter 2 we will design two amplifiers with this device: First, for the 1.9-GHz frequency band, having gain of 18 dB with 50-Ω input/output impedances. Then we will follow up with a *low-noise amplifier* (LNA) for the 1.9-GHz band.

1.7.1 Examining the effect of lossless feedback

The results of the broadband stability analysis plotted in Figure 1.36 looks scary at first. The device is potentially unstable ($\mu < 1.0$) up to 2.7 GHz,[11]

11. Most likely the device becomes stable at the low megahertz region, but data was not available at those frequencies.

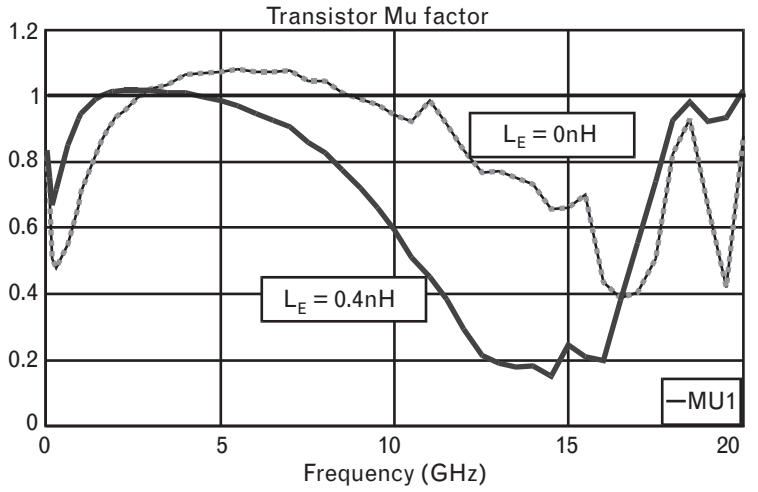


FIGURE 1.36
Effect of $L_E = 0.4$ nH series inductive feedback on stability and gain. Without the inductance the device shows a very high degree of potential instability in the low gigahertz frequency range, and also above 8 GHz. Adding the inductance improves the low-end stability, but degrades it for the higher frequencies.

and then again above 8.8 GHz. Such multiband instability generally requires multiple branch stabilizing networks instead of the single branches shown in Figure 1.35. Adding 0.4-nH inductance into the common (emitter) lead improves the low-end stability. The device is now stable from 1.7 to 4.9 GHz, although the high gigahertz range became worse. The 0.4-nH inductance here includes the unavoidable ground path, such as a via hole, and from now on it becomes part of the active device subcircuit.

As a general rule, the common terminal of an active device should be grounded directly. Since the small amount of inductance improves stability in the low gigahertz region, it will require less resistive loading to stabilize at those frequencies, leading to better dynamic range. Whether such feedback should be used or not depends on the intended operating frequency range and the *S*-parameters of the device. For example, if our target frequency is at 5.6 GHz, the feedback would make things worse for this device.

1.7.2 Device stabilization

From the *S*-parameters of the new two-port, we can compute the stability factor and the maximum gain, as shown in Table 1.4. The device becomes stable between 1.7 and 4.9 GHz, but remains potentially unstable below and above that frequency range.

Since potential instabilities exist in two separate frequency bands, we must use two stabilizing branches to cover the complete frequency range. First, let us handle the lower unstable frequency range that extends to 1.7 GHz. Then the resultant circuit will be stabilized for the high gigahertz range also.

Plotting the source and load stability circles of the subcircuit up to 3 GHz reveals the regions of unstable source and load terminations (Figures 1.37 and 1.38).

Table 1.4 Stability Factor and Gain of the Device with the Added 0.4-nH Emitter Inductance

Frequency (GHz)	μ_1	s_{21} (dB)	G_{MAX} /MSG (dB)
0.05	0.83	36.7	44.7
0.1	0.74	36	40.3
0.2	0.64	33.9	36.0
0.3	0.68	31.8	33.8
0.6	0.82	26.9	29.2
0.9	0.91	23.6	26.1
1.0	0.93	22.7	25.2
1.4	0.98	19.9	22.5
1.7	1.00	18.2	20.7
1.8	1.01	17.8	20
1.9	1.01	17.3	19.4
2.2	1.02	16.2	18.1
2.6	1.02	14.8	16.7
3	1.03	13.7	15.5

Note: *MSG* is the maximum stable gain a potentially unstable device can provide. G_{MAX} is the maximum gain of a stable device.

Looking at Figures 1.37 and 1.38, we can see that the output port is more convenient for stabilization since only a relatively small portion of the Smith chart represents terminations leading to possible oscillation. We used an admittance chart in Figure 1.38 because the unstable region is closer to the high-impedance (low admittance) portion of the chart. In such a case, adding a parallel stabilizing branch is more practical. The minimum necessary conductance is easy to determine from an admittance chart by locating the constant-conductance circle tangent to the unstable load region.

1.7.2.1 Parallel resistive stabilization for the low gigahertz frequency range

To stabilize the device in the illustrated frequency range (0.1–3.0 GHz), the minimum normalized parallel conductance is $g = 0.25$, equivalent to $r = 4.0$, $R = 200\Omega$. It is a good practice to create an added margin of stability by using a parallel resistor with 10% to 20% lower value. We will choose a 180-Ω resistor for our example. However, adding such a parallel resistor to the output port reduces the gain at all frequencies, even where the

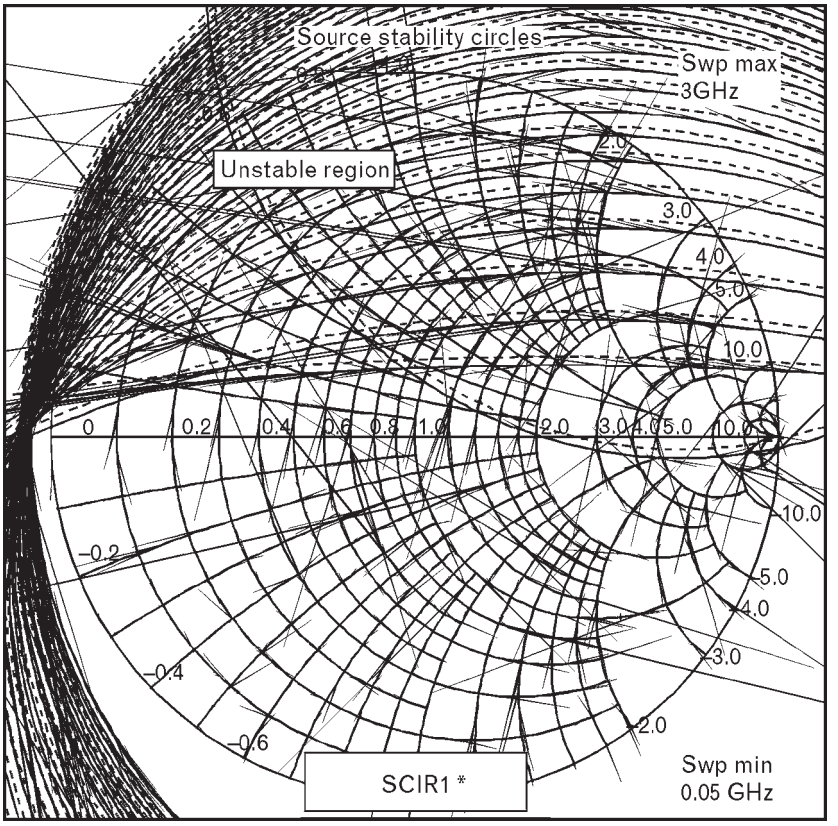


FIGURE 1.37
Source stability circles between 100 MHz and 3 GHz show that nearly half of the Smith chart represents unstable terminations. Since open circuit is in the unstable (dashed) region, series resistive stabilization is not possible. Short circuit is also very close to the unstable region; therefore, parallel resistance would not be practical. We conclude that the input side of the device is not a good one for stabilization.

device was already stable. If we choose instead a resistive-reactive branch [see Figure 1.35(b, c)], we can stabilize at the lower frequencies without sacrificing gain around 2 GHz.

In Chapter 2 we will show how to design various types of linear RF amplifiers, and one of the examples is a 1.9-GHz amplifier with simultaneous conjugate match. Let us stabilize our device for that task by adding the parallel branch containing a resistor and a short-circuited parallel transmission line stub of Figure 1.35(c). Our new circuit now has three components added to the transistor, as shown in Figure 1.39.

The parallel short-circuited stub presents an open circuit at its quarter-wavelength (90°) frequency, as well as others where the wavelength is an additional 180° (270°, 450°, and so forth). At those frequencies the impedance of the two-element branch is also infinite and the branch does not provide any stabilization. We set the parallel stub's electrical length to 90° at 2.0 GHz. Now the stabilizing branch has *finite impedance* at 1.9 GHz that further improves device stability.

Stability analysis of the circuit, depicted in Figure 1.40, proves the effectiveness of the added parallel branch. The device is now stable up to 8.7 GHz, with the exception of a glitch at 6 GHz where the parallel short-

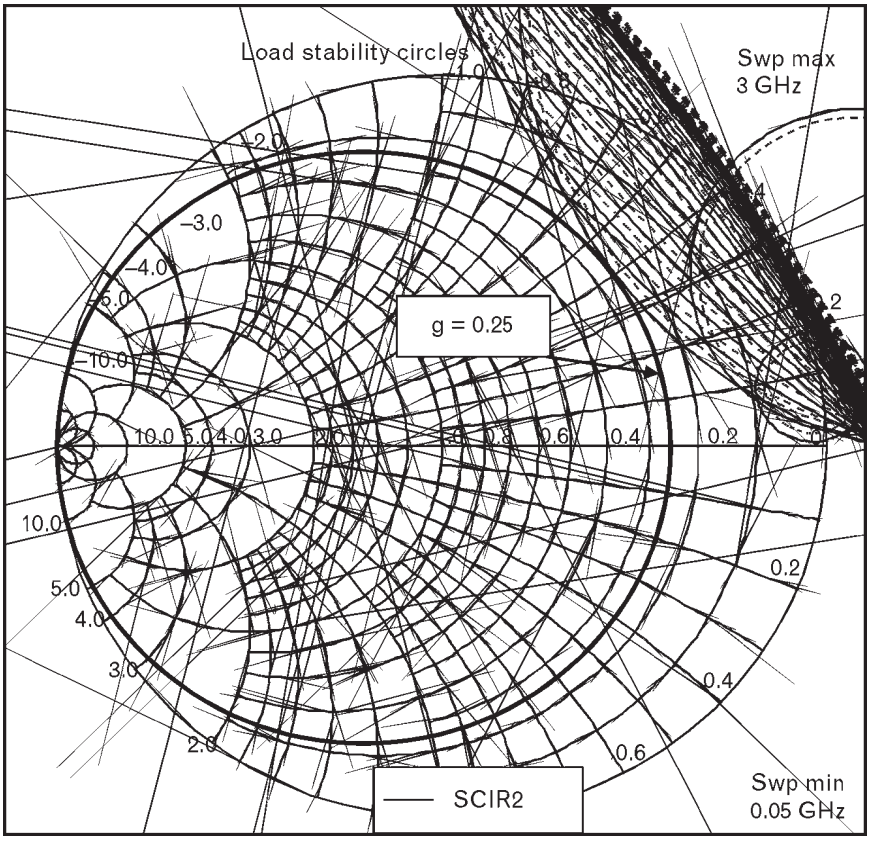


FIGURE 1.38
The unstable load region is much smaller than the unstable source region; therefore, it is easier to apply stabilization on the output side. The minimum parallel conductance, to stabilize for all frequencies, is $g_{PMIN} = 0.25$, since that normalized constant conductance circle is tangent to the stability circle that indicates the greatest instability.

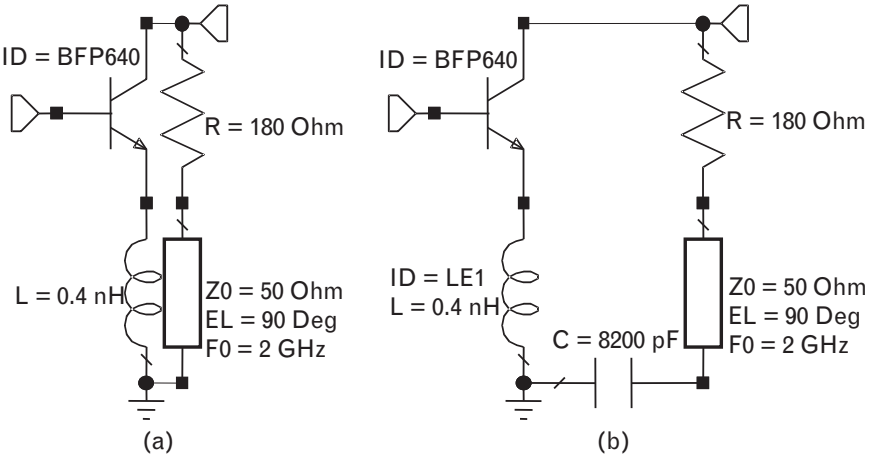


FIGURE 1.39
(a) Parallel stabilizing branch helps to stabilize to nearly 9 GHz. (b) Modified circuit to provide dc blocking capacitor for the parallel branch. Later when physical models will be used, the transmission line length will have to be shortened to compensate for the self-inductance of the 8200-pF capacitor.

circuited stub's wavelength is 270°. Unfortunately, the additional emitter inductance increased potential instability in the 9- to 18-GHz range, which also needs treatment. A second branch, however, will help us to accomplish that task.

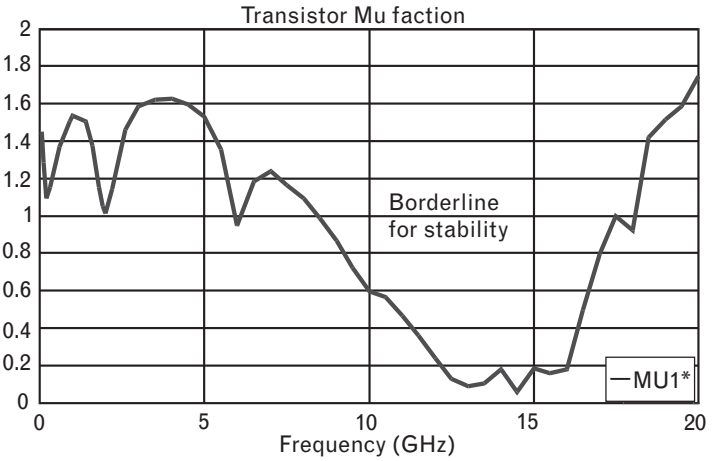


FIGURE 1.40
With the added parallel stabilizing branch, the low gigahertz region is stable, but instability is deepened in the 8- to 17-GHz range. A second branch is necessary to stabilize the device at those higher frequencies.

1.7.2.2 Series resistive stabilization for the high gigahertz range

To decide the location and configuration of the second stabilizing branch, we now plot the stability circles up to 20 GHz for the partially stabilized circuit of Figure 1.39(a). Once again, the source side of the device [see Figure 1.41(a)] is very difficult to work at because there are unstable regions at both high and low impedance regions of the Smith chart. The output side [Figure 1.41(b)], however, looks more promising. Here, an $r = 0.92$ unit normalized resistance (i.e., $R = 46\Omega$) gives the borderline stability. Increasing the series resistor to 50Ω provides an added margin of safety.

While selecting the stabilizing resistor values, we need to keep in mind that the resistors we select must behave like resistors through the entire frequency range. Of course, no ideal resistors exist, but compensated 50-Ω film resistors are available at a premium price and we could use one in the series stabilizing branch. The 180-Ω resistor of the parallel branch is much

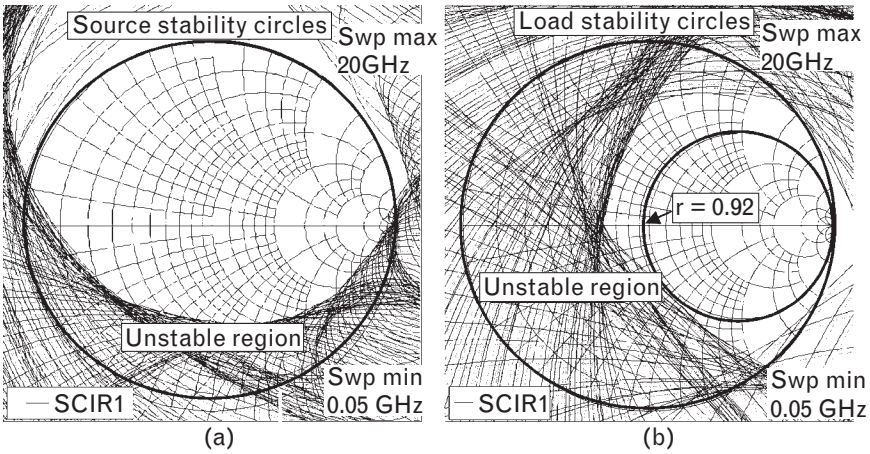


FIGURE 1.41
(a) Unstable source region is so widespread that broadband stabilization at the source side is very difficult. (b) The unstable region at the load side is mainly at low impedances, and an $r = 0.92$ unit normalized series resistance ($R = 46\Omega$) can be used to stabilize the device from 50 MHz to 20 GHz.

harder to realize and we will use two 91-Ω parts in series instead. (We show in Volume I, Chapter 7, that low-value resistors are inductive and high-value resistors are capacitive in the RF-MW frequency range. The crossover between these two extremes is around 90Ω to 100Ω where even low-cost uncompensated thick-film components are reasonably close to pure resistance.)

Once again, if we just add a good RF quality 50-Ω (standard value) series resistor at the output, it cuts the gain at all frequencies, including the low gigahertz range where the device was already stabilized. One possible solution is to bypass the resistor for our intended operating frequency range (i.e., in the vicinity of 1.9 GHz) by a resonant L-C circuit. However, since we will bypass the resistor for the 1.9-GHz band, we could again use a 91-Ω resistor, which is a less expensive option. The resonant parallel network shown in Figure 1.42 has very low impedance around 1.9 GHz, but the impedance quickly increases for the higher frequencies where the additional stabilization is needed.

While the circuit of Figure 1.42 looks simple, it may complicate dc biasing later[12] if we chose to bias the device through the output-matching network. (For narrow passbands, 5–10%, we can always add a parallel short-circuited quarter-wavelength stub for biasing. Beyond that, the presence of the stub reduces the effective bandwidth.)

We can modify the series stabilizing network leaving out the capacitor of the series resonant circuit, using only a parallel R-L network, letting the dc bias flow through the inductor, as shown in Figure 1.43. While this approach adds about 0.5 dB more loss in the 1,900-MHz range, it can cover broader bandwidths and later simplify dc biasing. An additional benefit is that this slightly higher loss leads to increased stability around 1,900 MHz.

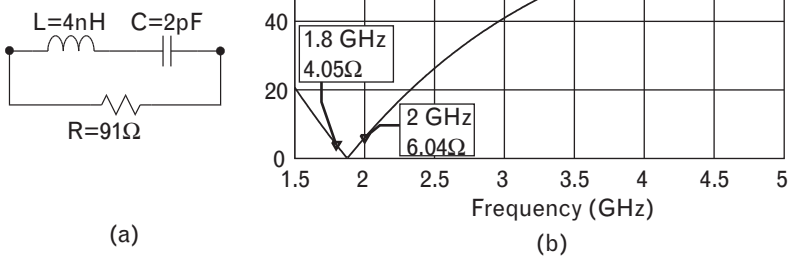


FIGURE 1.42
(a) Adding a simple series L-C circuit across the stabilizing resistor causes a short circuit at (b) resonant frequency and reasonably low impedance through the 1.8- to 2.0-GHz range.

12. It is generally a good practice to have free dc access to the transistor terminals, without any element that either blocks dc or causes dissipative loss.

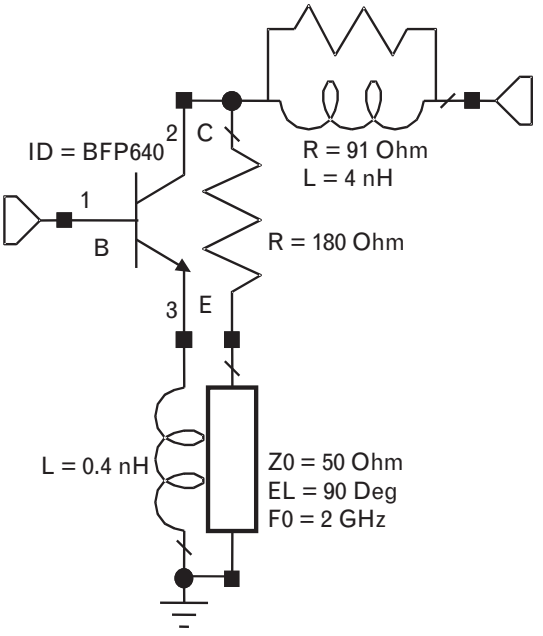


FIGURE 1.43
RF schematics of the final stabilized active device circuit using ideal circuit elements.

1.7.2.3 Broadband stability analysis

After the addition of this second stabilizing branch, the new active circuit (Figure 1.43) is unconditionally stable up to 20 GHz (the range for which we have measured transistor data). Using this stabilized device eliminates the possibility of unexpected oscillation, as long as the source and load terminations are passive and there are no external feedback paths between the two ports.

Results of a broadband stability analysis (Figure 1.44) verify unconditional stability at all frequencies, showing the cyclical effect of the parallel

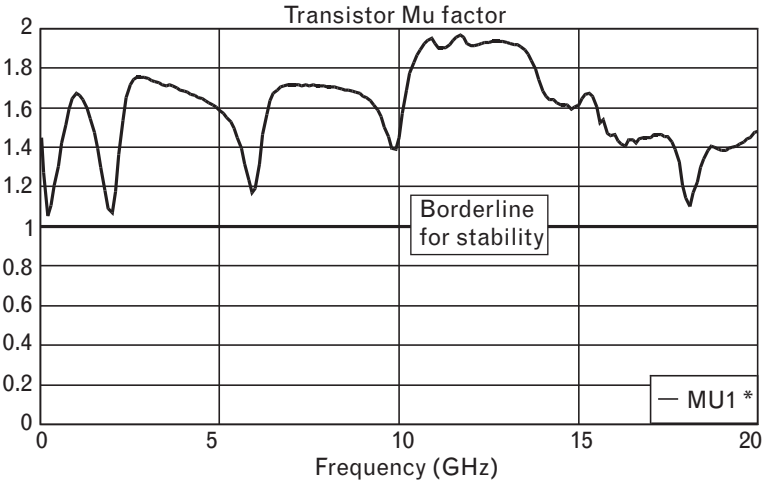


FIGURE 1.44
Broadband stability analysis of the stabilized device shows unconditional stability ($\mu > 1.0$) through the 0.1- to 20-GHz range. Stabilization was achieved by three circuit segments: an emitter inductor, a parallel branch, and a series branch at the output. A new data set called BFP

branch. Since the quarter-wave frequency of the short-circuited parallel stub was set to 2 GHz, the parallel branch loses its stabilizing effect also at 6, 10, 14, and 18 GHz as well. At those frequencies, the μ-factor is closer to unity value. Still, considering that matching circuit losses will most likely improve stability, we do not have to worry about oscillation. (We will address the physical component related issues in Section 1.10.)

Figure 1.45 shows that the unstable regions of the source and load stability circles are now outside of the Smith chart (thick circle trace), since the device is stable for all frequencies. For better visibility, we use here the compressed chart with radius of 3.0.

With the stabilized device we can now design amplifiers at 1.9 GHz and expect gain over 18 dB, and over 20 dB in the 900-MHz band (see Table 1.5), assuming both ports are simultaneously matched. We can now create a new two-port data set for the stabilized device to be used in Chapter 2.

Although the concepts we covered here are valid for all cases, our goal is to use the device later in the 1.9-GHz amplifier design illustration of the next chapter. Since the device has sufficient gain, we could afford to be conservative and sacrifice about 5-dB gain to ensure broadband stability. Selection of device stabilization should always be specific to the operating frequency range and component specifications.

We should also point out again that our stabilization exercise so far has been based on *ideal circuit elements*. Real-life components have parasitics and possible multiple resonances, as we discuss in Volume I, Chapter 7. Circuit layout also affects RF performance, and these factors need to be considered as well. It is very likely that when those effects are included, the component values and perhaps even the topologies will change. Although a completely detailed simulation of the physical circuit is beyond the scope of this book, in Section 1.10 we will look at the response of the stabilized device with some physical component models.

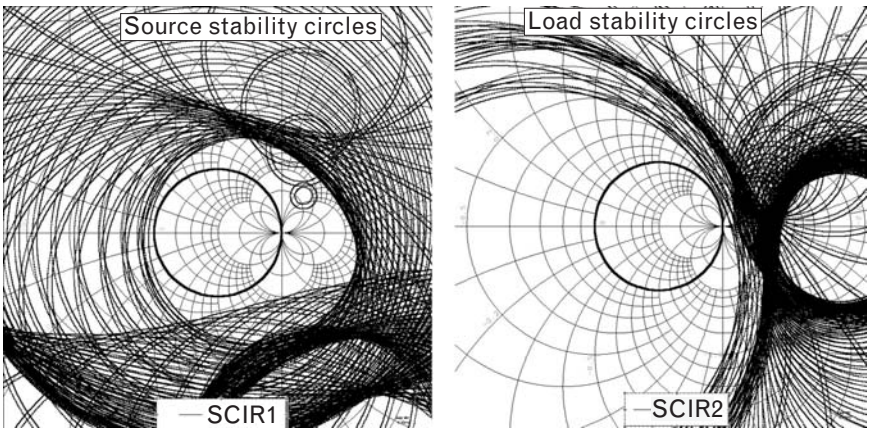


FIGURE 1.45
The compressed Smith chart (radius = 3.0) allows seeing results beyond the conventional unit-radius chart. In this example, all the stability circles either completely enclose or they are outside the unit-radius chart.
Either way, the full unit-radius chart refers to stable terminations

Table 1.5 Stability μ-Factor and Maximum Decibel Gain of the Stabilized Device Shown in Figure 1.43

Frequency (GHz)	μ_1	s_{21} (dB)	G_{MAX} (dB)
0.05	1.45	31.2	36.4
0.1	1.27	30.8	35.7
0.2	1.05	29.7	34.5
0.3	1.11	28.4	31.8
0.6	1.42	24.4	25.9
0.9	1.65	21.3	22.3
1.0	1.67	20.4	21.3
1.4	1.55	17.6	18.5
1.6	1.4	16.7	17.8
1.8	1.18	16.4	17.9
1.9	1.09	16.4	18.2
2	1.07	16.3	18.0
2.2	1.36	14.7	15.5
2.6	1.74	11.8	12.3
3	1.75	10.0	10.6

Note: Although we show only the data up to 3 GHz, the device is unconditionally stable for all frequencies. An interesting observation is that in the 2.6- to 3.0-GHz range the maximum gain is virtually the same as the 50-Ω gain, indicating that the input and output impedances of the device are close to 50Ω.

RF stability analysis should be a vitally important part of any active circuit and system design. (Remember that an ounce of caution may prevent a pound of sorrow!) We will continue the stability discussion in Chapters 4 and 6 with nonlinear device models.

1.8 The dc bias techniques

Since active devices require dc bias to operate, we also need to design additional circuitry for that function. The importance of dc bias circuitry is frequently underestimated even though it is crucial to successful RF operation. RF parameters of transistors vary with the changes in dc bias, as shown in Figure 1.46. Obviously if the bias conditions of the device change, the RF performance will also shift.

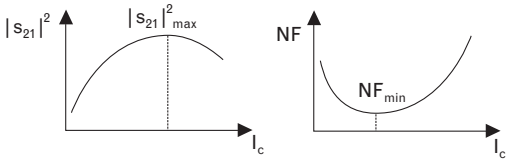


FIGURE 1.46 *Typical variation of bipolar transistor gain and noise figure as a function of dc bias current. Note that maximum gain and minimum noise occur at different bias currents.*

Tolerance variations of transistor dc parameters are much larger than those of RF parameters. For example, the magnitude of the forward transmission coefficient, $|s_{21}|$, of an RF transistor may vary ±20% from one production lot to another. The $|s_{11}|$ and $|s_{22}|$ generally have less than ±10% deviation in magnitude. In comparison, some of the critical dc parameters of a bipolar transistor, such as h_{FE}, might vary as much as 200% to 300%. When we also consider the temperature dependency of h_{FE}, V_{BE}, and bias resistors, the total dc bias change is even more significant [21]. A well-designed dc bias circuit must be able to compensate for the effects of such large variations. Therefore, we do not exaggerate by stating that the design of the bias circuit may be as important as the RF circuit design. Unfortunately, RF designers frequently do not invest time in a thorough dc circuit selection and simulation—and they pay the price later when the RF performance fails.

1.8.1 Passive dc bias networks

It is a good practice to use some form of feedback [22, 23] in the bias circuit to minimize the dc voltage and current variations of the device. There are several forms of possible negative feedback circuit configurations, and some of them are shown in Figure 1.47. All of these options are dissipative (i.e., they take power away from the dc source). When the power loss is critical, we need to consider active bias circuitry, covered in Section 1.8.2. Active biasing offers a higher level of dc stability.

Since a common-emitter configuration gives a 180° phase change between collector and base at dc, any resistive connection between those terminals provides negative feedback. For example, in the circuits of Figure 1.47(a–c), any increase of collector current increases the voltage drop across the collector resistor and thereby lowers the collector voltage of the device. A lower voltage difference between collector and base also reduces the base voltage and current and cuts back the collector current.

We prefer to dc ground the RF transistor directly instead of adding a bias resistor into the common lead, even though the emitter feedback (or source feedback for FETs) is a very effective technique for dc bias stability, as shown in Figure 1.48. At RF frequencies, bypassing a bias resistor in

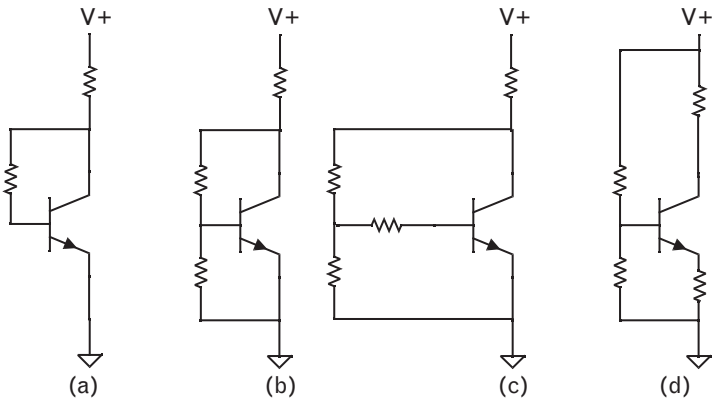


FIGURE 1.47 *Various resistive negative feedback bias circuits for bipolar transistors in the order of increasing effectiveness: (a) collector-base parallel feedback only, (b) collector-base parallel feedback with voltage divider, (c) collector-base parallel feedback with voltage divider plus current source resistor, and (d) emitter feedback.*

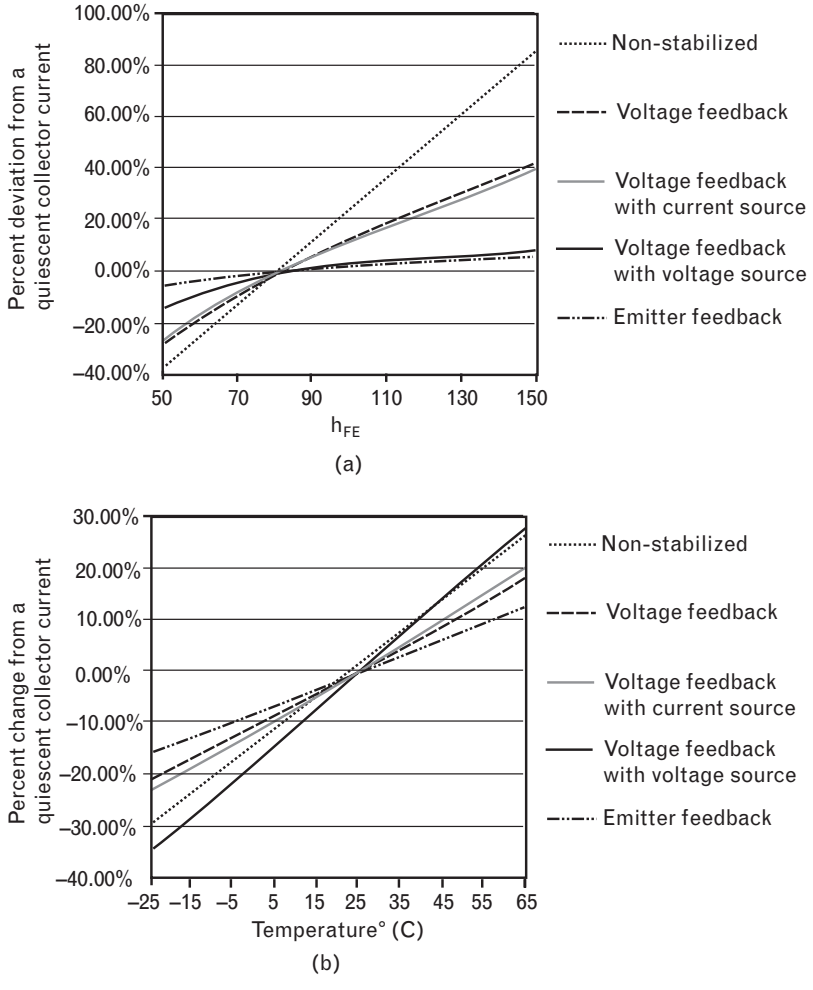


FIGURE 1.48 *Percent change of collector current versus (a) hFE variation and (b) temperature change, showing the importance of dc stabilization. In low-voltage, high-current applications active bias control is necessary. (Source: [23]. © 2003 Agilent Technologies. Reprinted with permission.)*

the common ground [like Figure 1.47(d) and Figure 1.49(c)] is not easy due to component parasitics and resonances, particularly in broadband applications.

Since there are two fundamental modes of operations [24, 25] for FETs—depletion (normally "ON" with zero gate bias) and enhancement (normally "OFF" with zero gate bias) modes—we treat them differently. Biasing depletion-mode FETs is more difficult since they require dual-voltage dc supplies to keep the gate negative with respect to the source. If the dual supply is not available, resistive voltage drop may be used to create the negative bias for the gate. Figure 1.49 shows some of the optional bias configurations for depletion-mode devices, but of the three circuits, only one [Figure 1.49(c)] incorporates negative feedback through self-biasing. In this last method the source resistor, R_S, must be bypassed for the appropriate RF range, and the circuit also needs a higher supply voltage to overcome the voltage drop caused by R_S.

Enhancement-mode FETs operate similarly to bipolar transistors since the gate is forward biased. The source terminal can be directly grounded, which is always desirable at RF. Passive and active bipolar bias circuits shown in Figure 1.47(c) and Figure 1.50(a) (with positive gate voltage) are also applicable for enhancement-mode devices. Since excessive forward

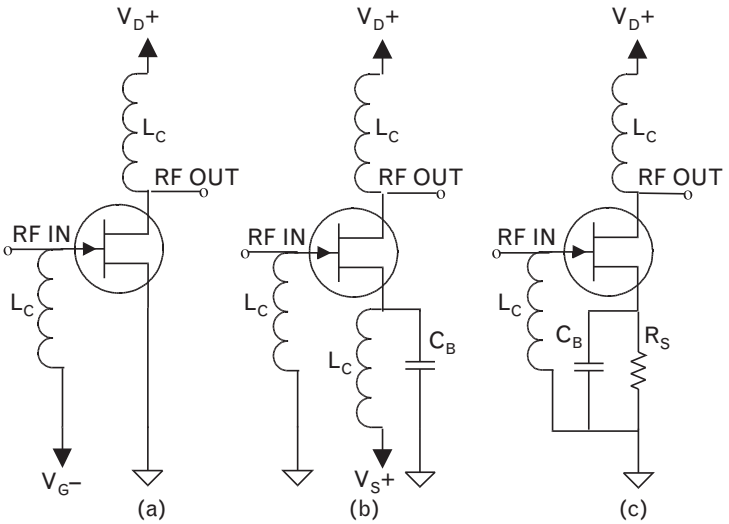


FIGURE 1.49 *The dc bias arrangements for FETs: (a) with positive (V_D) and negative (V) supplies referenced to ground, (b) dual positive supply (V_D and V_S), and (c) with single positive supply, the current through R_S sets the gate voltage negative with respect to the source of the FET. Capacitors (C_B) represent RF shorts,[13] and inductors (L_C) are bias chokes.[14] In circuits (b) and (c) the bypass capacitor may cause stability problems.*

13. A very low impedance capacitor to present RF ground.

14. Inductor with very high impedance at RF.

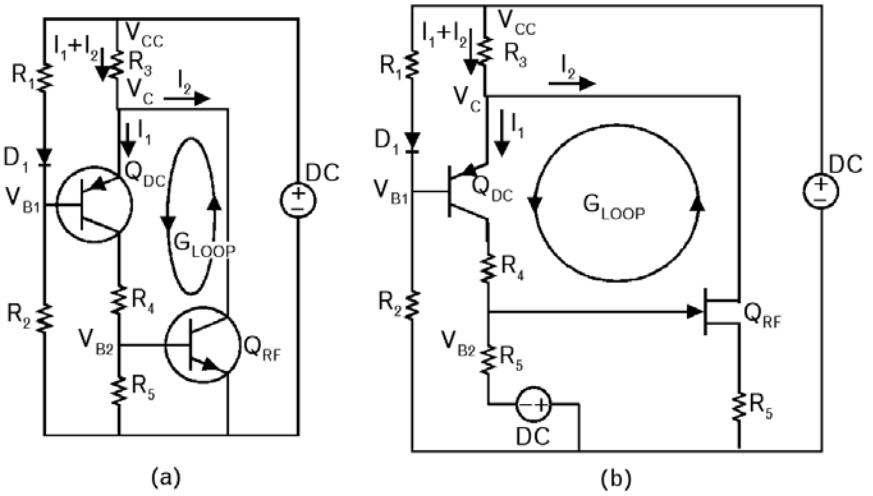


FIGURE 1.50 *Active dc bias circuits for (a) bipolar transistor and enhancement mode FETs and (b) depletion-mode FETs. Q_{dc} is the bias transistor and Q_{RF} is the RF device. At dc such circuits offer negative feedback and a high degree of bias stability. The dc feedback loop's gain (G_{LOOP}) must be carefully filtered to avoid low-frequency oscillation.*

bias can easily damage the device, *always* use a series protective resistor [26] in the gate circuit.

1.8.2 Active dc bias circuits

Since dc feedback always lowers the power supply voltage available, it may be difficult to have effective feedback considering today's requirement for low voltage operation. If feedback is not practical or sufficient and statistical analysis shows significant bias-circuit variation, perhaps some active biasing is necessary. Active biasing might be achieved by a special function circuit or by adding another low-frequency transistor that controls the dc bias voltage. Figure 1.50 shows two possible dc circuit topologies for active bias applications: one for bipolar [27] and one for a field effect transistor. In both cases we must isolate the active bias transistor from the RF device through a broad frequency range. RF isolation is needed to avoid losses in the bias transistor. Equally important is low-frequency isolation to avoid low-frequency oscillation (sometimes called *motor-boating*) within the feedback loop formed by the two devices.

In Figure 1.50(a), the base voltage, V_{B1}, of the bias transistor is set at about 0.75V to 0.8V below V_C, the desired collector voltage of Q_{RF}. The

current through resistor R_3 is set by the voltage difference between V_{CC} and V_C, forming a constant current source of $(I_1 + I_2)$. Most of the total current flows through Q_{RF}, since I_2 of Q_{RF} is much larger than I_1 of Q_{dc}. The diode between R_1 and R_2 offsets the temperature dependency of the base-emitter junction of Q_{dc}.

For a depletion-mode RF FET the active circuit is similar [Figure 1.50(b)], but now we need a dual power supply to apply negative bias to the gate of the FET.

In low-voltage circuit application the active bias schemes of Figure 1.50 may cause an excessive voltage drop. An alternative approach is to use commercially available bias controllers [28, 29] that may require as little as 0.1V drop of the dc supply. Integrated into a single small package that also includes all related control circuitry, these components offer high dc gain and unconditional stability.

1.8.3 Feeding dc bias into the RF circuit

Once the dc bias scheme is selected, the next step is to apply the voltages and currents to the active RF devices. There are three possible ways to do this:

1. In low-current applications, it is practical to connect the bias resistors directly to the RF terminals of the transistor. In this case we need to do a thorough modeling of the bias network because the resistors may have significant parasitics that affect RF operation.

2. Apply the dc bias through RF chokes. This approach is quite practical for narrowband operation, but not at all useful for broadband applications. A word of advice: *Be sure that the bias choke is truly high impedance at the operating frequency range.* A typical mistake is to choose a large value inductor (say, in the millihenry range) and expect it to be high impedance at RF frequencies. Unfortunately, due to internal parasitics, a large inductor resonates at the low megahertz range, and in the gigahertz range it has totally different behavior. It may not even behave as an inductor any more, as we see in Volume I, Chapter 7, under RF Component Models. Therefore, much smaller inductors (50–300 nH) should be used that self-resonate in the center of the operating frequency range. Figure 1.51 illustrates this point by comparing the impedances of two inductors through the RF range. The 270-nH inductor resonates at 906 MHz and also acts as a virtual open circuit through a ±50-MHz band. A 3,000 times larger inductor, 820 μH, resonates around 3 MHz, and at 906 MHz it behaves more like a 120-Ω resistor rather than an inductor. Figure 1.51(b) also shows the

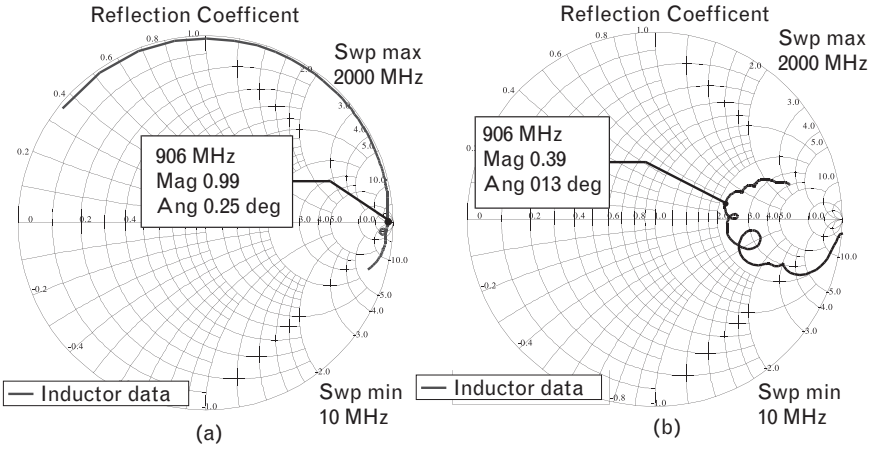


FIGURE 1.51
Reflection coefficients of two chip inductors: (a) 270 nH and (b) 820 μH. The (b) plot shows that larger inductance does not guarantee higher terminal impedance. At 906 MHz the 270 nH behaves nearly as an open circuit, while the 820-μH inductor represents much lower

higher-order resonances at frequencies above the primary parallel resonance.

3. Feed dc bias through matching elements to the transistor. This is the most desirable but the most difficult form to design because it only works with certain topologies. Circuits that use a parallel inductor, or a parallel short-circuited stub in the matching network, are appropriate for this task because we can feed the dc bias through these shunt elements as shown in Figure 1.52. In this case the bias network has absolutely no effect on the RF operation.

1.8.4 The dc bias circuit simulation

To provide an accurate dc analysis we must have a nonlinear model for the device. Most manufacturers readily provide SPICE-type models that can be used by circuit simulators that accept those models. Having dc

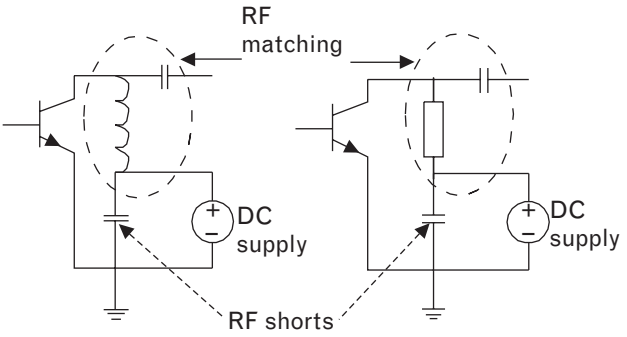


FIGURE 1.52 *Lumped and distributed matching networks that are also suitable for dc biasing. The parallel inductor and stub pass the dc current to the active device while the series capacitor serves as a dc block. Both circuits are highpass types, causing more attenuation at low frequencies. RF shorts are created by self-resonant capacitors at the center of a narrow passband.*

simulation capability enables us to look at the transfer characteristics of the device from which we can determine the required base current for the desired collector current and collector voltage setting. Once we know the base current, we can also measure the corresponding base-emitter voltage. With that information we can proceed with the dc bias network design. After deciding what bias circuit topology to use, since we have all voltage and current information, the element values are computed by simply using Ohm's Law—back to basics.

We should point out a frequently overlooked difference between dc-h_{FE} and ac-h_{FE}, sometimes referred to as h_{FE0}, of a transistor that can cause errors in bias calculations. The dc-h_{FE} is defined as

$$h_{FE(dc)} = \frac{I_C}{I_B}$$

The ac-β is given as

$$h_{FE(ac)} = \frac{\Delta I_C}{\Delta I_B}$$

The two circuits of Figure 1.53 are convenient for creating the transfer characteristics of the device and also for finding the voltage of the base-emitter junction for a specified base current.

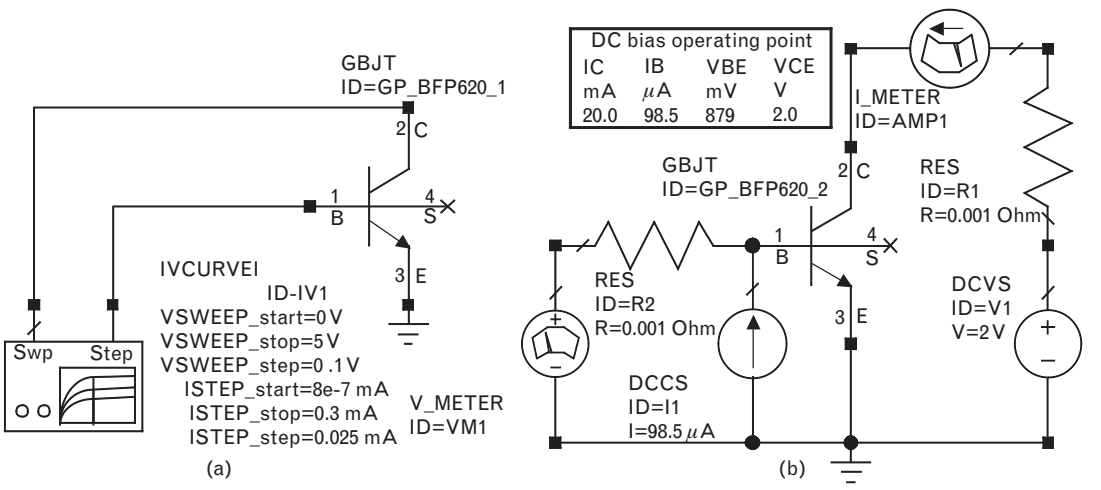


FIGURE 1.53 *Establishing the dc transfer characteristics and finding the dc bias parameters: (a) curve tracer setup in the MW Office simulator; and (b) driving 0.0985-mA current into the base (determined from I-V curves) results in 20-mA collector current. The corresponding base-emitter voltage is 879 mV.*

Figure 1.54 shows the transfer characteristics of the BFP 640 transistor. From the transfer characteristics we can now locate the necessary base current to have 2-V collector voltage and 20-mA collector current through the device. From the base current versus base voltage plot we can find out what kind of quiescent base voltage we need from our bias network.

To illustrate the resistive dc bias network design, we choose one of the resistive feedback networks shown in Figure 1.47, redrawn here as Figure 1.55. Since we will later use this device with *S*-parameters measured at V_{CE} = 2.0V and I_C = 20 mA, setting our supply voltage to 3.0V, we allow a 1.0-V drop across the resistor R_3. This voltage drop leads to 22-mW dc power dissipation in resistor R_3, which is one-third of the total dc power used by the amplifier. In portable applications, where power efficiency is very important, an active bias circuitry may be a better solution.

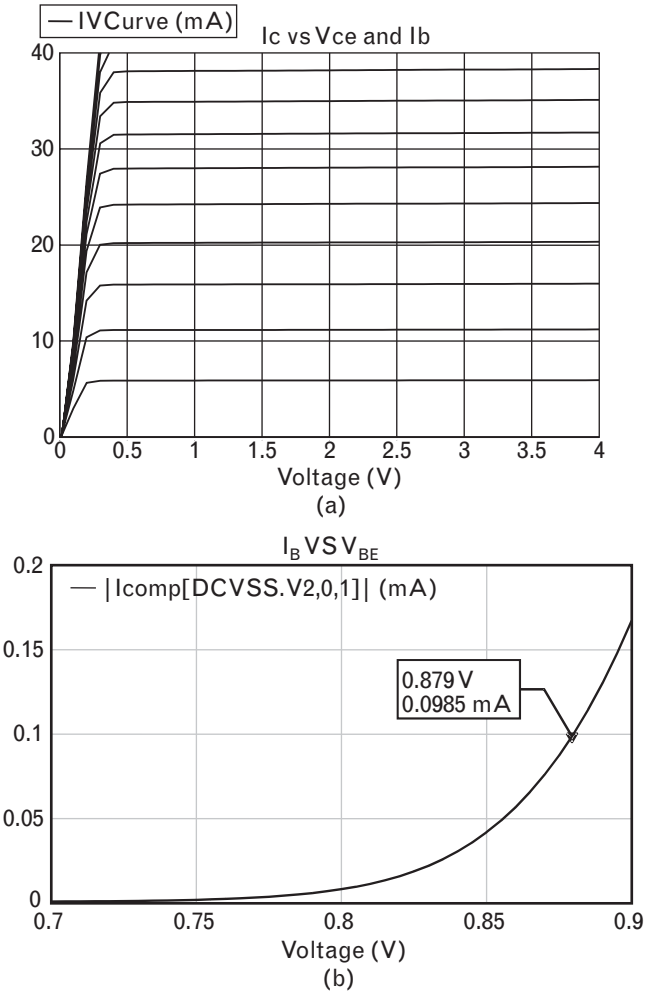


FIGURE 1.54
(a) The dc transfer characteristics of the BFP 640 transistor. For a desired quiescent bias point at V_{CE} = 2V, I_C = 20 mA, the corresponding base current (shown in milliamperes at the right side of the plot) is just under 0.1 mA.

(b) Plotting I_B versus V_{BE} at V_{CE} = 2V helps to determine the exact base voltage for I_B = 0.0985 mA, that is, 98.5 μA.

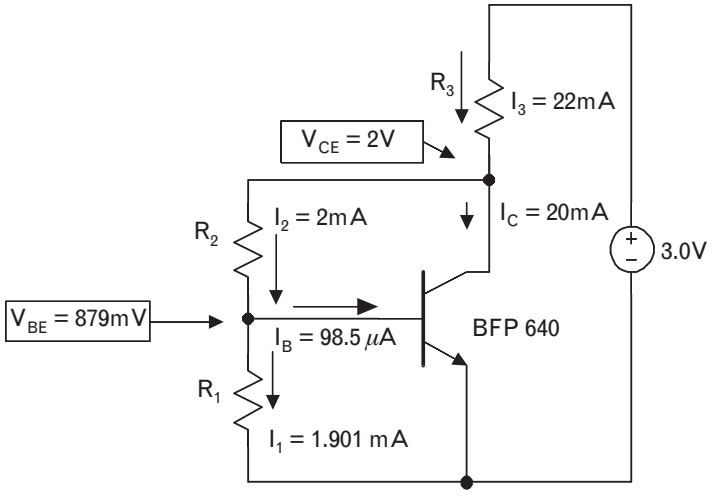


FIGURE 1.55
Resistive feedback circuit for dc biasing. The resistors must either be outside of the RF signal path or their true physical models must be included in the RF circuit simulation.

For dc stability, it is a good practice to run about 5% to 10% of the collector current through the resistive base-voltage divider. Knowing the required collector bias current is 20 mA, we can set the total divider current to 2.0 mA. Next, let us calculate the three resistor values of the bias circuit.

$$R_1 = \frac{879\,\text{mV}}{1.901\,\text{mA}} = 462\Omega$$

$$R_2 = \frac{(2{,}000 - 879)\,\text{mV}}{2\,\text{mA}} = 560\Omega$$

$$R_3 = \frac{(3-2)\text{V}}{0.022\,\text{A}} = 45.5\Omega$$

Simulating the dc operation of the circuit of Figure 1.55 with the true nonlinear model verifies the calculated collector voltage and current values.

An alternative and much quicker way to obtain the element values for the bias network is to use a general purpose engineering tool, such as the Agilent AppCAD program that can be downloaded for free from Agilent Technologies' Web site [30]. AppCAD allows us to choose one of several available bias network configurations, including the one we used in our example. The program calculates the resistor values for a specified V_{CE}, I_C, and V_{CC} combination (Figure 1.56). AppCAD provides the option to select the nearest available standard values using ±1%, ±2%, or ±5% component tolerances. The shortcoming of this program is that it uses a generic transistor model instead of a specific device. Therefore, when the exact device model is available and you have access to a nonlinear simulator, we encourage a full dc simulation with that model.

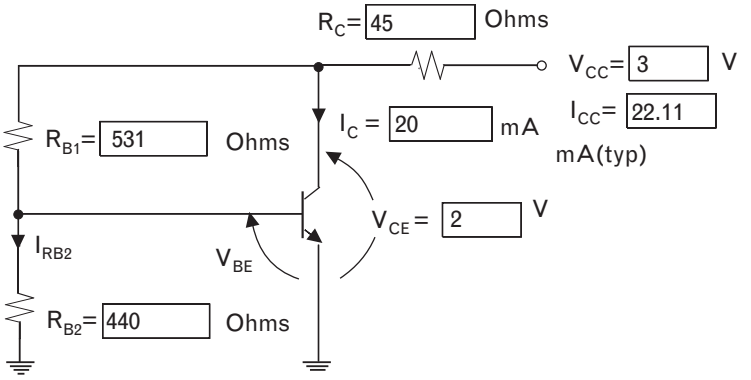


FIGURE 1.56 *AppCAD solution for the dc bias circuit of Figure 1.55 first gave us different results than our computations, due to the generic nonlinear model the program uses for the transistor. The most significant deviation was in the V_{BE} assumption. AppCAD's default was 780 mV (not shown here), while the actual value is 879 mV. Overriding the default V_{BE} input gave resistor values close to the computed values shown above.*

1.8.5 Filtering of dc bias networks

Improper filtering of the bias network can easily cause potential instability and even oscillation. In active bias circuits, such as in Figure 1.50, instability occurs generally at low frequencies because the bias transistor is a low-frequency device with no gain at RF. The loop marked G_{LOOP} forms negative feedback at dc. However, the loop gain can easily be 40 to 50 dB at some low frequency in the kilohertz range where the phase angle represents positive feedback. In some cases the loop may even have sufficient gain in the megahertz region to cause problems, necessitating careful analysis with true physical models or measured data for the components used for filtering.

Figure 1.57 shows the low-frequency loop-gain filtering of the active bias circuit of Figure 1.50(a). Two bias chokes (L_C) and two bypass capacitors (C_{BP}) form a lowpass filter to attenuate gain at frequencies where the loop's phase angle causes positive feedback.

1.9 Statistical and worst-case analyses

The dc-bias circuit design is not completed until a thorough statistical and/or worst-case analysis is performed with satisfactory results. It is imperative to find out how much the nominal operating bias point shifts through extreme temperature and dc variations in beta. We mentioned those two parameters because those are the most critical, but of course if we had additional available tolerances for the transistor, we should apply all known parameters. Remember that the *S*-parameters, as well as other RF parameters, are all bias dependent. If the bias conditions change

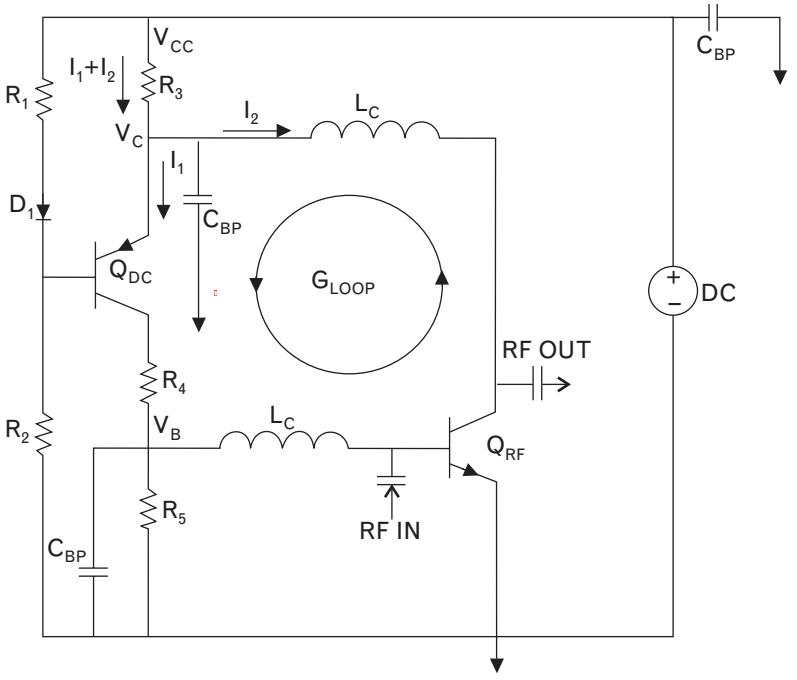


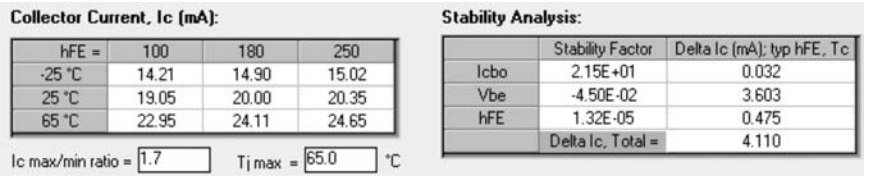
FIGURE 1.57
Separating the RF signal from the dc bias loop, G_{LOOP}. The loop filter elements, L_C and C_{BP}, must be carefully chosen to prevent low-frequency instability without significantly affecting RF amplification.

significantly (i.e., more than 5–10%), the performance will also change, and we need to be aware of the variations.

AppCAD also provides worst-case collector currents for extreme h_{FE} and temperature variations. Again, the calculations are based on the generic model, but if no other information is available the AppCAD results give us a good idea of the worst-case performance, as shown in Figure 1.58. Collector current drops at low temperature and minimum h_{FE} combination. Highest collector current is reached when the temperature and h_{FE} are both at their maximums.

The quick worst-case analysis shows the effect of negative feedback in the dc-bias circuit. For an h_{FE} variation of 100 to 250 through the –25°C to +65°C temperature range, the collector current changes about ±25%. Still, that much dc-bias variation may reduce the RF performance to an unacceptable level. This circuit may require an active bias arrangement.

A complete statistical or worst-case analysis requires nonlinear tempe ature-dependent models for the transistors as well as temperature coefficients for the resistors of the dc bias circuit. In discrete RF circuits the RF stages are generally dc blocked from each other; therefore, the accuracy of the active device models are not so critical—in most cases we can get reasonable results by using generic models. However, in integrated circuits where a large number of transistors are directly connected together, the modeling task is crucial. In such cases we must find and use the appropriate nonlinear models for all active devices.



Collector Current, Ic (mA):

hFE =	100	180	250
-25 °C	14.21	14.90	15.02
25 °C	19.05	20.00	20.35
65 °C	22.95	24.11	24.65

Ic max/min ratio = 1.7 Tj max = 65.0 °C

Stability Analysis:

	Stability Factor	Delta Ic (mA); typ hFE, Tc
Icbo	2.15E+01	0.032
Vbe	-4.50E-02	3.603
hFE	1.32E-05	0.475
	Delta Ic, Total =	4.110

FIGURE 1.58 *Worst-case analysis of the dc bias circuit of Figure 1.55 not including temperature effects of the bias resistors. The collector current varies by a ratio of 1.7 through a temperature range of −25°C to + 65°C with h_{FE} changes between 100 and 250. Active bias arrangement can reduce the extreme current variation significantly, at the cost of increased circuit*

1.10 Circuit layout considerations

Last but not least, we need to discuss the effect of circuit layout on RF performance. Unfortunately, many designers do not want to or simply cannot have control over the layout, and it is passed over to a PC board group that has very little training or understanding of RF principles. Circuit board layout departments primarily have one consideration: Pack as much circuitry as possible into the smallest available space. Therefore, they cannot be blamed that the RF circuit behaves completely different from the initial simulated performance.

Most of the latest RF/MW simulators allow us to translate our circuit schematic into an actual layout. Even if that layout is not acceptable by the company's PC board group, because it may not pass all the design rules, at least it gives them general guidance for how the circuit should be laid out for RF considerations. It is very important at this point to develop a good communication link between the design engineer and the board layout group. Otherwise, lengthy and expensive prototype cycles will delay the project completion. Having a close relationship between design and PC board departments also helps the design engineers appreciate the limitations of what can and cannot be realized under realistic production conditions.

EM simulators can be a great help in making critical decisions about circuit board layout, particularly when multilayer boards are used. In such cases the RF signals may pass through various layers using via holes or other connection techniques. The effect of these interconnecting links must always be thoroughly analyzed when frequencies reach the gigahertz range.[15] For example, the parallel stabilizing branch shown in Figure 1.39(b) uses a transmission line with quarter-wave length at 2 GHz. The physical length of this transmission line may be too long to place on the top layer, and it may be more practical to place it on one of the inside layers of a multilayer PC board. The lower node of the transmission line stub is most likely RF grounded by a suitable capacitor, instead of getting a direct short circuit

15. The same applies to high-speed digital circuits, covered in Volume I, Chapter 9.

to ground. Since the capacitor cannot be placed inside the PC board, the lower node of the stub must be brought back to the top of the board where the capacitor is placed. The lower side of the capacitor, of course, must then be grounded, which means going through another via hole again to the layer where the ground is located (Figure 1.59). Obviously the true effective electrical length of the transmission line is altered by these additional components and only an EM simulator can give us proper results.

Figure 1.60 shows a possible layout for the parallel stabilizing branch, using a four-layer PC board. The dc bias and RF circuitries are frequently placed on different layers to minimize circuit size and maximize isolation. The three surface-mount components, (91 + 91)-Ω resistors and 8200-pF capacitor, are placed on the top layer. Two buried via holes (V2-V5) make contact with the third conductor layer where the quarter-wave transmission line stub is located. Via hole V1 grounds the 8,200-pF capacitor to the second conductor, which is a ground layer.

In our illustrations, via holes do not completely pass through the PC boards, which is an expensive manufacturing operation, as we cover in Volume I, Chapter 7. A less expensive and more commonly use form is when the plated via holes pass through all layers of the board and partially behave like open-circuited stubs to the rest of the network [31].

When circuits are realized with multilayer PC board technology, circuit simulators can only provide approximate solutions. A complete 3-D, or even 2.5-D, EM simulation, although desirable, may not be practical for everyone. Circuit designers need to combine best available models and computer-aided design tools with engineering and economic judgment at this point of the design cycle.

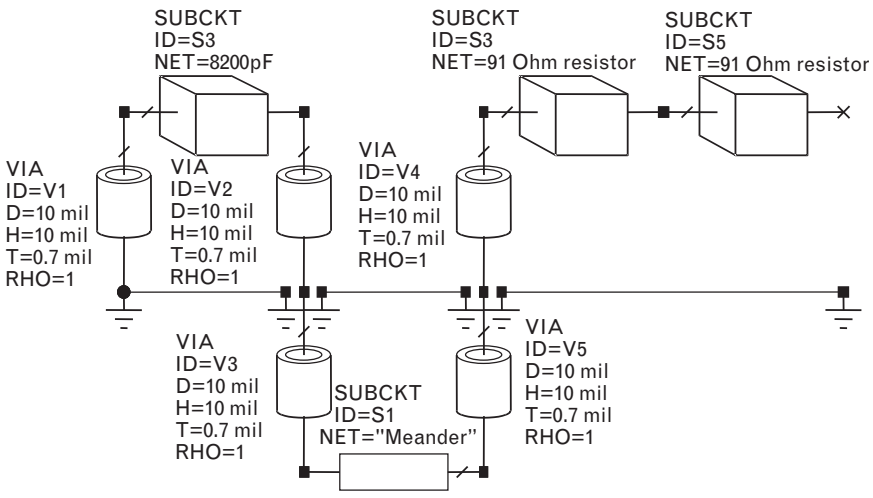
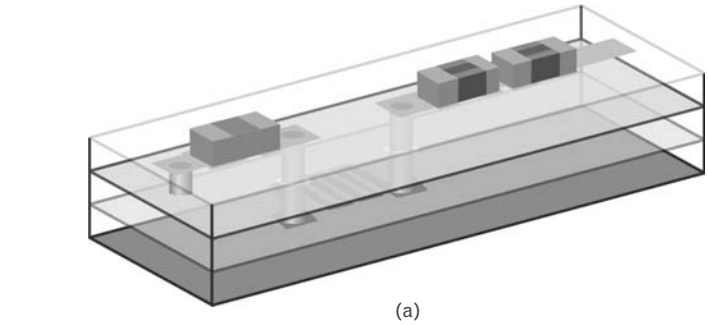


FIGURE 1.59
The parallel transmission line branch of Figure 1.39(b) is realized on three different layers of a five-layer PC board.



(a)

(b)

FIGURE 1.60
(a) The three-dimensional artistic recreation of the PC board shows different types of via holes. Two of the vias pass through the ground layer, and the shorter one grounds the chip capacitor. The second and fourth layers from the top are RF ground. (b) The top view shows the component placement on the top of the PC board.

Figure 1.61 compares the cosimulated[16] stability of the ideal circuit of Figure 1.43 with a second form using the true physical stabilizing circuit to 8 GHz—the upper limit of available measured data for the passive components. Capacitor data was provided by Murata (GRM36X7R682K50) and the resistors were characterized by Modelithics (RRNA91). The 180-Ω resistor was modeled by a series connection of two 91-Ω parts. The 4-nH inductance was realized as a small loop on the PC board. Interestingly, the physical circuit is more stable, which is most likely caused by the greater losses of physical components with the increase of frequency.

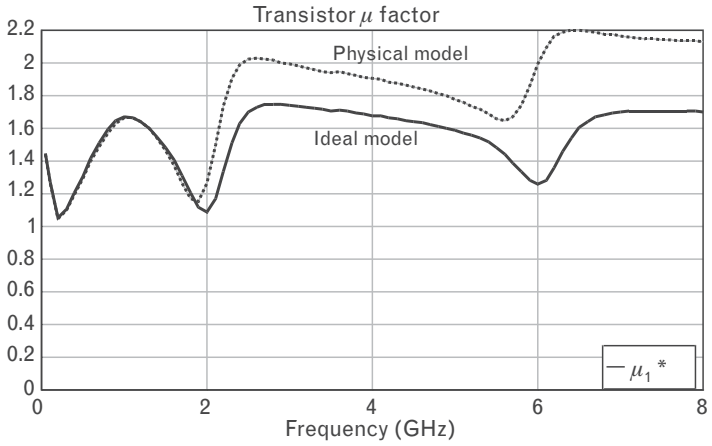


FIGURE 1.61
Comparison of the stabilized BFP 640 device with ideal and real physical component models in the stabilizing branches. Measured data for the passive components was only available to 8 GHz. EM simulation was performed with Sonnet's EM 2.5-D program.

16. Combining circuit and EM simulations.

1.11 Summary

Active RF circuit design is a challenging task that requires a thorough understanding of RF fundamentals, active and passive component models, ac/dc design, and CAD techniques, including EM simulations and also some understanding of device physics. RF and dc circuit stability analysis and circuit layout are just as important as the design of the RF matching networks. Successful product design calls for a combination of theory and practical skills that may take years to develop. Although modern CAD tools are very helpful, they still do not replace sound engineering judgment.

1.12 Problems

For all listed problems, download from http://www.infineon.com the broadband two-port *S*-parameters and noise-parameters of the Infineon BFP 405 transistor at 2-V, 2-mA bias condition. Use any available RF circuit simulator to perform the calculations. The AppCAD program is available through http://www.agilent.com.

1. Design an amplifier stage for G_{UMAX} with the BFP 405 device at 880 MHz, without any added stabilization, using ideal lumped matching elements. What are the gain, input and output reflection coefficient magnitudes of the amplifier with (a) $|s_{12}|$ set to zero and (b) using the actual s_{12} of the device? How does the value of G_{UMAX} compare with the computed *MSG* of the device?
2. Design a *resistive* stabilizing network for the BFP 405 to have $\mu_1 = 1.05$ between 0.01 and 6 GHz. If necessary, use more than a single branch. Find out what is G_{MAX} at 880 MHz after the stabilizing network was added. Modify the stabilized network(s) by adding reactive element(s) to improve the maximum gain at 880 MHz without giving up unconditional stability for all frequencies.
3. Design a resistive dc bias network for the BFP 405 using $V_{CE} = 2$V, $I_C = 2$ mA, $V_{BE} = 0.85$V, and $I_B = 25\ \mu$A. Use a 3.0-V dc supply and assume a dc h_{FE} of 80 at 25°C. Check your results with the resistor values computed by the AppCAD. What kind of worst-case collector current variations does AppCAD predict for h_{FE} changes between 50 and 150 through temperature range of –25°C to +65°C?
4. Extra credit, requiring access to a nonlinear circuit simulator: Download the SPICE model of the BFP 405 and design an active dc bias network shown in Figure 1.50(a) with the specification of

Problem 3. Use any generic PNP model for the bias transistor. Compare the collector current variation of the BFP 405 with the results of Problem 3 through the same worst-case conditions.

References

[1] Application Note 95, "*S*-Parameters, Circuit Analysis and Design," Hewlett-Packard, Palo Alto, CA, September 1968.

[2] Application Note 154, "*S*-Parameter Design," Hewlett-Packard, Palo Alto, CA, April 1972.

[3] Besser, L., "A Fast Computer Routine to Design High Frequency Circuits," *IEEE ICC Conference Digest*, San Francisco, CA, June 1970

[4] Bandler, J. W., and C. Charalambous, "A New Approach to the Computer-Aided Design of Microwave Circuits," *IEEE MTT Int. Symp. Digest*, Arlington Heights, IL, 1972.

[5] Bandler, J. W., et al., "A Microwave Network Optimization Program," *IEEE MTT Int. Symp. Digest*, Boulder, CO, 1973.

[6] Besser, L., et al., "Computer Aided Design for the 1980s," *IEEE MTT Int. Symp. Digest*, Los Angeles, CA, 1981.

[7] Gonzalez, G., *Microwave Transistor Amplifiers Analysis and Design*, 2nd ed., Englewood Cliffs, NJ: Prentice Hall, 1997.

[8] Grivet, P., *Microwave Circuits and Amplifiers*, Paris, France: Academic Press, 1974 (French), London, England: Academic Press, 1976 (English).

[9] Moschytz, G., *Linear Integrated Networks*, New York: Van Nostrand Reinhold, 1974.

[10] Vendelin, G. D., M. Pavio, and U. L. Rhode, *Microwave Circuit Design Using Linear and Nonlinear Techniques*, 2nd ed., New York: Wiley Interscience, 2003.

[11] Losee, F., *RF Systems, Components, and Circuits Handbook*, Norwood, MA: Artech House, 1997.

[12] Bennett, W. R., *Electrical Noise*, New York: McGraw-Hill, 1960.

[13] Linvill, J. G., and J. F. Gibbons, *Transistors and Active Circuits*, New York: McGraw-Hill, 1991.

[14] Woods, D., "Reappraisal of the Unconditional Stability Criteria for Active Two-Port Networks in Term of *S*-Parameters," *IEEE Trans. on Circuits and Systems*, February 1976.

[15] Edwards, M. L., and J. H. Sinsky, "A New Criterion for Linear Two-Port Stability Using a Single Geometrically Derived Parameter," *IEEE Trans. on Microwave Theory and Techniques*, December 1992.

[16] Mason, S. J., "Feedback Theory—Further Properties of Signal Flow Graphs," *Proc. of IRE*, July 1956.

[17] MacLean, D. J., "Stability Margins in Microwave Amplifiers," *IEEE Trans. on Microwave Theory and Techniques*, March 1984.

[18] Wang, K., M. Jones, and S. Nelson, "The *S*-Probe: A New Cost-Effective Method for Evaluating Multi-Stage Amplifier Stability," *IEEE MTT International Symposium Digest*, 1992.

[19] Besser, L., "Stability Considerations of Low-Noise Amplifiers with Simultaneous Noise and Power Match," *IEEE MTT Int. Symposium Digest*, May 1975.

[20] Vendelin, G. D., "Feedback Effects on GaAs MESFET Performance," *IEEE MTT Int. Symposium Digest*, May 1975.

[21] Sze, S. M., *Physics of Semiconductor Devices*, 2nd ed., New York: Wiley, 1981.

[22] Application Note AN 944-1, "Microwave Transistor Bias Circuit Considerations," Agilent Technologies, 1976.

[23] Agilent Technologies, "A Comparison of Various Bipolar Transistor Biasing Circuits,"Agilent Technologies, 2003.

[24] Soares, R., (ed.), *GaAs MESFET Circuit Design*, Norwood, MA: Artech House, 1988.

[25] Pengelly, R., *Microwave Field-Effect Transistors*, 3rd ed., Atlanta, GA: Noble Publishing, 1995.

[26] Application Note 1222, "High Intercept Low Noise Amplifier Using the Agilent AFT-54143 Enhancement Mode PHEMT," Agilent Technologies, 2002.

[27] Vizmuller, P., *RF Design Guide, Systems, Circuits, and Equations*, Norwood, MA: Artech House, 1995.

[28] Application Note 014, "Application Considerations for the Integrated Bias Controller, BCR400W," Infineon Technologies, Munich, Germany, 2002.

[29] Application Note 064, "Using the BCR410W Bias Controller with RF Transistors," Infineon Technologies, Munich, Germany, 2002.

[30] http://www.agilent.com.

[31] Swanson, Jr., D. G., *The Field-Solver Circuit Design Cookbook,* Norwood, MA: Artech House, 2003.

SELECTED BIBLIOGRAPHY

Hardy, J., *High Frequency Circuit Design*, Englewood Cliffs, NJ: Reston/Prentice Hall, 1979.

Kennedy, G., *Electronic Communication Systems*, 3rd ed., New York: McGraw-Hill, 1985.

Krauss, H. L., C. W. Bostian, and F. H. Raab, *Solid State Radio Engineering*, New York: John Wiley & Sons, 1980.

Kurokawa, K., *An Introduction to the Theory of Microwave Circuits*, New York: Academic Press, 1969.

Ludwig, R., and P. Bretchko, *RF Circuit Design, Theory, and Applications*, Englewood Cliffs, NJ: Prentice Hall, 2000.

Rollett, J. M., "Stability and Power Gain Invariants of Linear Two-Ports," *IRE Trans. on CT*, March 1962.

Stern, A. P., "Stability and Power Gain of Tuned Transistor Amplifiers," *Proc. of IRE,* March 1957.

CHAPTER 2

Linear and low-noise RF amplifiers

Amplifier applications may require minimum noise, maximum gain, maximum power output, best impedance matching, stability into varying loads, wide bandwidth, cascading with other circuits, and other performance factors. This chapter presents practical linear amplifier design techniques that are used to meet the different performance requirements.

2.1 Introduction

Linear RF amplifiers fulfill various tasks in communication systems, and most of them can be grouped into the following three major categories:

- *Low noise:* At the input of a receiver the signal level may be very low. In addition to amplifying the signal, we must exercise special care to minimize the unavoidable noise contribution of the amplifier. The source termination (i.e., an antenna) may vary during normal operation, and the amplifier must function in spite of the changes. The output may see highly reactive terminations outside the passband, presented by filters that follow the amplifier, and it must be stable for all those terminations. Low noise considerations and the active device noise parameters are covered in Section 2.5, using the *available gain* technique.
- *Maximum small-signal gain:* After the signal is raised well above the noise level, gain becomes a more important factor than noise. Also, since the amplifier may face a wide range of terminations at various frequencies, RF stability is another key consideration. These intermediate level amplifiers are designed for maximum gain, with simultaneously matched input and output ports. We discuss these amplifiers in Section 2.2 with the *transducer gain* approach.
- *Maximum absolute output power:* At the output of a transmitter, high power level is the major concern, although in wireless systems linearity and efficiency are often just as important. Load termination change is also another important consideration in mobile transmitter applications. Linear power amplifier design methodology is part of

Section 2.4, using the *operating gain* technique. Chapters 3 and 5 are dedicated to the details of nonlinear device modeling and high-power amplifier design.

There are also special amplifier applications, functions such as *limiting* and *isolation* (or buffer), but in this chapter we focus on the above three categories. To keep the length of the chapter reasonable, we again rely on cited references for derivations and proofs and use mathematical expressions only when absolutely needed to show the underlying principles.

2.2 Bilateral RF amplifier design for maximum small-signal gain

The beauty of the *S*-parameter amplifier design approach [1] lies in its simplicity. We characterize the active two-port with measured *S*-parameters instead of a complex equivalent circuit model. Then, we find two terminations that satisfy our performance requirements.

Simple enough? Yes. Can it be used for all types of active component applications? Unfortunately, no, since *S*-parameters are functions of dc bias, operating temperature, and applied signal level. In addition, *S*-parameters apply to steady-state conditions only. Some system components, like oscillators, mixers, or class-C amplifiers, require time-variant nonlinear design techniques using nonlinear models for the active devices (see Chapters 3–7).

Restating the transducer power gain, G_T, (1.2) as (2.1),

$$G_T = \frac{\left(1-\left|\Gamma_S\right|^2\right)\left|s_{21}\right|^2\left(1-\left|\Gamma_L\right|^2\right)}{\left|\left(1-s_{11}\Gamma_S\right)\left(1-s_{22}\Gamma_L\right)-s_{12}s_{21}\Gamma_S\Gamma_L\right|^2} \tag{2.1}$$

we see that G_T is a function of the source and load terminations (Γ_S and Γ_L) and of the *S*-parameters of the two-port shown in Figure 2.1. If we know all those parameters, the gain computation is quite straightforward. Designing an amplifier for a specific gain with given *S*-parameters, however, is more complicated, because we have one equation with two unknowns. In Section 1.4.2 we handled cases where one of the two terminations was fixed, and (1.2) was reduced to have a single unknown. To find two terminations simultaneously, we must have two equations.

If the amplifier is to produce the maximum small-signal power gain available from the active device, we must find a unique solution for two terminations to impedance-match both ports simultaneously. Naming those two terminations as $\Gamma_S = \Gamma_{MS}$ and $\Gamma_L = \Gamma_{ML}$, we can write two equations to specify simultaneous conjugate match at both ports:

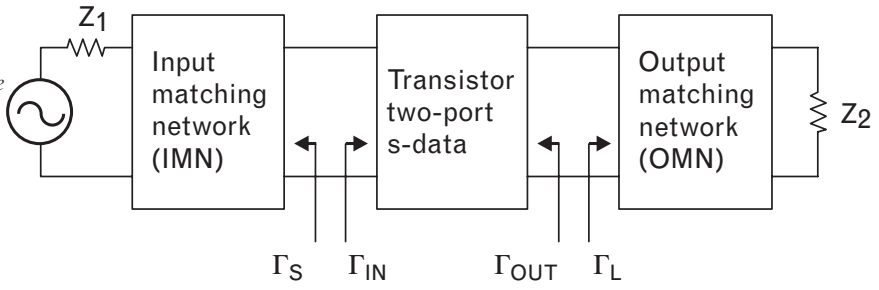


FIGURE 2.1
Block diagram of an RF amplifier where the active device is characterized by measured two-port S-parameters. Performance is a function of the applied terminations, Γ_S and Γ_L.

$$\Gamma_{MS} = \Gamma_{IN}^{\star} \tag{2.2}$$

and

$$\Gamma_{ML} = \Gamma_{OUT}^{\star} \tag{2.3}$$

Γ_{MS} is the unknown unique source termination to match whatever Γ_{IN} it sees at the input port while the output is terminated with Γ_{ML}. Similarly, Γ_{ML} is the unknown unique load termination to match the resultant Γ_{OUT} that it sees at the output port while the input is terminated with Γ_{MS}.

Substituting (1.3) into (2.2) using $\Gamma_L = \Gamma_{ML}$ gives

$$\Gamma_{MS} = \left(s_{11} + \frac{s_{12}\, s_{21}\, \Gamma_{ML}}{1 - s_{22}\, \Gamma_{ML}} \right)^{\star} \tag{2.4}$$

and similarly, substituting (1.5) into (2.3) using $\Gamma_S = \Gamma_{MS}$ gives

$$\Gamma_{ML} = \left(s_{22} + \frac{s_{12}\, s_{21}\, \Gamma_{MS}}{1 - s_{11}\, \Gamma_{MS}} \right)^{\star} \tag{2.5}$$

Solving (2.4) and (2.5) for the two unknowns, Γ_{MS} and Γ_{ML}, gives [2]

$$\Gamma_{MS} = \frac{B_1 - \sqrt{B_1^2 - 4|C_1|^2}}{2C_1} \tag{2.6}$$

and

$$\Gamma_{ML} = \frac{B_2 - \sqrt{B_2^2 - 4|C_2|^2}}{2C_2} \tag{2.7}$$

where

$$B_1 = 1 + |s_{11}|^2 - |s_{22}|^2 - |\Delta|^2 \quad C_1 = s_{11} - s_{22}^{*}\Delta \quad |\Delta| = |s_{11}s_{22} - s_{12}s_{21}|$$
$$B_2 = 1 + |s_{22}|^2 - |s_{11}|^2 - |\Delta|^2 \quad C_2 = s_{22} - s_{11}^{*}\Delta$$

Equations (2.6) and (2.7) are valid for all unconditionally stable two-ports. This approach provides physically realizable passive terminations when the active two-port is unconditionally stable, or has already been stabilized by the addition of an appropriate external stabilizing network. For academic interest, we point out that in a potentially unstable two-port with $K < 1$ there are no simultaneous conjugate match solutions, and in a potentially unstable two-port with $K > 1$ and $|\Delta| > 1$ (a case that does not occur in practical designs), the solutions of Γ_{MS} and Γ_{ML} are given by (2.6) and (2.7) with a plus sign in front of the radical. Furthermore, for such a case the resulting simultaneous conjugate matched transducer power gain is a *minimum* [2].

Let us emphasize that our bilateral design computations of Γ_{MS} and Γ_{ML} are exact—they include the effect of input-output interaction. If we transform the existing source and load impedances to Γ_{MS} and Γ_{ML}, and apply the new terminations to the unconditionally stable two-port, the overall circuit is matched at both ports and we realize the simultaneously conjugate matched maximum gain, called G_{MAX} (see Figure 2.2).

Substituting Γ_{MS} for Γ_S and Γ_{ML} for Γ_L in (2.1) leads to a solution for G_{MAX} of the two-port as

$$G_{MAX} = \frac{\left(1 - |\Gamma_{MS}|^2\right)|s_{21}|^2\left(1 - |\Gamma_{ML}|^2\right)}{\left|\left(1 - s_{11}\Gamma_{MS}\right)\left(1 - s_{22}\Gamma_{ML}\right) - s_{21}s_{12}\Gamma_{MS}\Gamma_{ML}\right|^2} \tag{2.8}$$

Considering the complexity of Γ_{MS} and Γ_{ML}, the full form of (2.8) is much longer than the width of this page. It can, however, be simplified [3] to a more practical form:

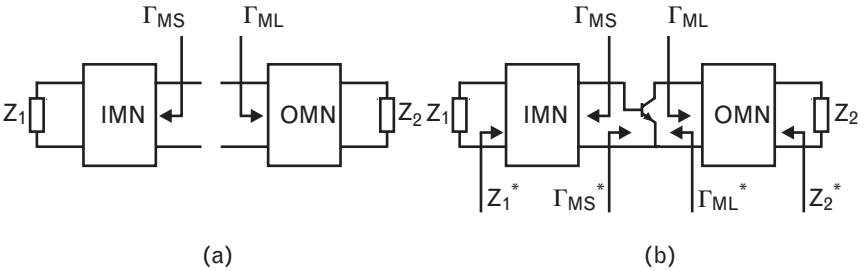


FIGURE 2.2 *(a) In maximum gain amplifiers the actual source and load terminations Z_1 and Z_2, are transformed to Γ_{MS} and Γ_{ML}. (b) Placing the unconditionally stable two-port between Γ_{MS} and Γ_{ML} matches the amplifier to Z_1 and Z_2.*

$$G_{MAX} = \left|\frac{s_{21}}{s_{12}}\right|\left(K - \sqrt{K^2 - 1}\right) \tag{2.9}$$

where K is the frequency-dependent stability factor of the two-port, defined in Section 1.5.3.1. Equation (2.9) is valid for unconditionally stable two-ports only, because if $K < 1$, we get a negative quantity inside the radical.

For a potentially unstable device, we define the *maximum stable gain* (*MSG*), which is the highest theoretically realizable gain with passive terminations, after the device is stabilized with cascaded resistance to borderline stability, that is, to achieve $K = 1$.

$$MSG = \left|\frac{s_{21}}{s_{12}}\right| \tag{2.10}$$

Transistor data sheets often combine G_{MAX} and *MSG* under one common heading. They show *MSG* at frequencies where the device is potentially unstable and G_{MAX} at other frequencies. As mentioned before, the stability factor is rarely listed.

We should consider a word of caution about *MSG*. To build a circuit with an active device at borderline stability, the required source and load terminations are located at the circumference of the Smith chart.[1] We cannot create real-life impedance matching circuits for such requirements. In practice, we must slightly "overstabilize" the device to get physically realizable matching networks. Doing this reduces the maximum gain to about 1 or 2 dB less than the computed value of *MSG*. By the time we include the effects of component losses, the *practical maximum stable gain* (*PMSG*) is 2 to 3 dB less than the advertised value.

When feedback is applied to partially or fully stabilize a device, the original *MSG* may be reduced by several decibels, as shown in the following illustration. We first use the data sheet *S*-parameters of the BFP 640, without any stabilization (see Table 2.1), and compute *MSG* in decibels. Then, we compare *MSG* with G_{MAX} of the stabilized device.

Computing *MSG* from the above *S*-parameters at 1.9 GHz gives us

$$\text{MSG}_{\text{dB}} = 10\log\left|\frac{s_{21}}{s_{12}}\right| = 10\log\left(\frac{10.7}{0.05}\right) = 23.3\,\text{dB}$$

As we will see in Table 2.2, G_{MAX} at 1.9 GHz is 18.2 dB and is 5.1 dB less than the initial MSG of the device. There are two reasons for the reduction: first, we used inductive series feedback to reduce the region of

1. If the device is potentially unstable, Γ_{MS} and Γ_{ML} have magnitudes greater than unity.

Table 2.1 Common Emitter *S*-Parameters of the BFP 640 at 1,900 MHz, Biased at 2V and 20 mA

Frequency (GHz)	s_{11} Mag	Ang	s_{21} Mag	Ang	s_{12} Mag	Ang	s_{22} Mag	Ang
1.9	0.28	152	10.7	81	0.050	55	0.35	-46

Note: Without stabilization the device has basic transducer power gain of $10\log |s_{21}|^2 = 20.6$ dB. Angles (Ang) are given in degrees.

Table 2.2 Stabilized *S*-Parameters of the Infineon BFP 640 Device Biased at 2V, 20 mA

Frequency (GHz)	s_{11}		s_{21}		s_{12}		s_{22}	
	Mag	Ang	Mag	Ang	Mag	Ang	Mag	Ang
1.8	0.28	–82	6.63	76	0.060	70	0.49	4.8
1.9	0.25	–92	6.64	72	0.066	67	0.51	–1.7
2.0	0.20	–92	6.51	66	0.071	62	0.49	–9.6

Frequency (GHz)	μ-factor	s_{21} (dB)	G_{MAX} (dB)	Γ_{MS}		G_{ML}	
				Mag	Ang	Mag	Ang
1.8	1.18	16.4	17.9	0.28	145	0.55	6.42
1.9	1.09	16.4	18.2	0.41	166	0.65	13.6
2.0	1.07	16.3	18.0	0.46	–179	0.68	20.5

Note: Since the μ-factor is greater than unity, the device is unconditionally stable. If this stable two-port is simultaneously terminated with Γ_{MS} and Γ_{ML}, the gain at 1.9 GHz is $G_{MAX} = 18.2$ dB.

unstable terminations on the Smith chart. Then, we added resistive stabilizing branches at the output that raised the μ-factor above borderline stability. At 1.9 GHz we have sufficient gain to make this sacrifice. At higher frequencies, however, we need to be careful to maintain sufficient gain for amplification.

2.2.1 Illustrative exercise: amplifier design for maximum gain, G_{MAX}

Now that we know how to compute the necessary terminations for maximum gain, let us design an amplifier with the BFP 640 device, already stabilized in Section 1.7. We target the 1.85- to 1.95-GHz band by designing the matching circuits at 1.9 GHz with 50-Ω system terminations. *S*-parameters and various computed RF parameters of the stabilized device are listed in Table 2.2.

We can make an interesting observation about the computed G_{MAX} values in Table 2.2. The stabilized device has more gain at 2.0 GHz than at 1.8 GHz. The reason is that at 2 GHz the device is less stable because the parallel stabilizing branch of Figure 1.39 behaves as an open circuit at the quarter-wave frequency of the short-circuited stub. At 2 GHz that parallel branch does not help the stability of the device. Above 2 GHz the stub again represents finite impedance, and stability begins to improve until the stub reaches its half-wavelength frequency at 4 GHz. At that point the branch provides maximum stability. The cycle repeats periodically every 4 GHz from there on.

Choosing the matching network topologies is the next step. For our first attempt, we use a single-section highpass configuration that offers two advantages:

- The unwanted low-frequency gain is rolled off. If higher selectivity and/or better frequency response symmetry is needed, we can later increase the order of the networks.

- When appropriate, a series capacitor/parallel short-circuited stub combination is very convenient for feeding through and blocking dc bias.

Figure 2.3 shows the input and output matching network determinations on the Smith chart. For both circuits we start from the existing 50-Ω system terminations ($Z_0 = 50\Omega$), and with highpass matching sections, transform to Γ_{MS} and Γ_{ML} at 1.9 GHz. Since bilateral analysis includes the effects of input-output interaction, our results are exact. Assuming lossless matching components, the computed maximum gain is 18.2 dB.

The location of Γ_{ML} is ideal for choosing a two-element matching circuit that is also convenient for dc biasing. Of course, we need to dc isolate the ground terminal of the parallel stub. A capacitor self-resonant at 1.9 GHz is used for an RF ground, as shown in Figure 2.4.

Reading the normalized reactance and susceptance magnitudes of the two output port matching elements from the Smith chart (Figure 2.3) gives

$$x_{CO} = 1.92 \text{ and } b_{SSO} = 0.52$$

Computing the output matching network's element values from the reactance and susceptance, using the normalized formulas from Volume I, Chapter 2 (also summarized in the appendix of this volume),

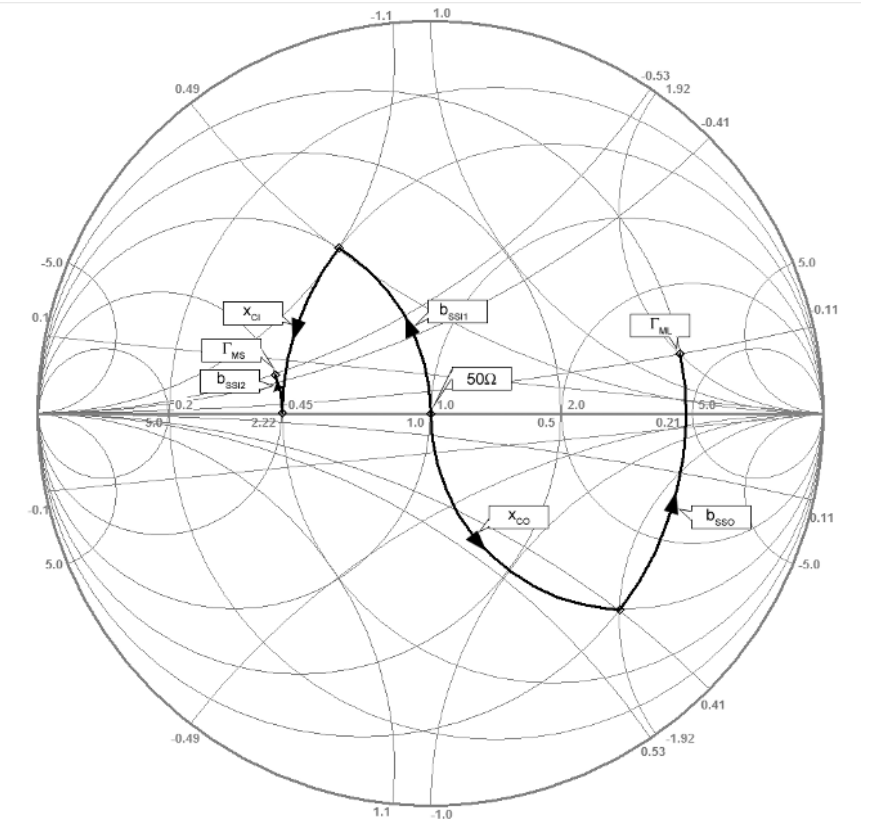


FIGURE 2.3 *Finding the simultaneous conjugate matched source and load terminations of the stabilized BFP 640 at 1.9 GHz. Γ_{MS} and Γ_{ML} are the terminations the device needs to see at the input and output for maximum gain, G_{MAX}. Input and output networks are transformed to Γ_{MS} and Γ_{ML} from the 50-Ω terminations. With slight modifications, the three-element input network and the two-element output network are also suitable for dc biasing.*

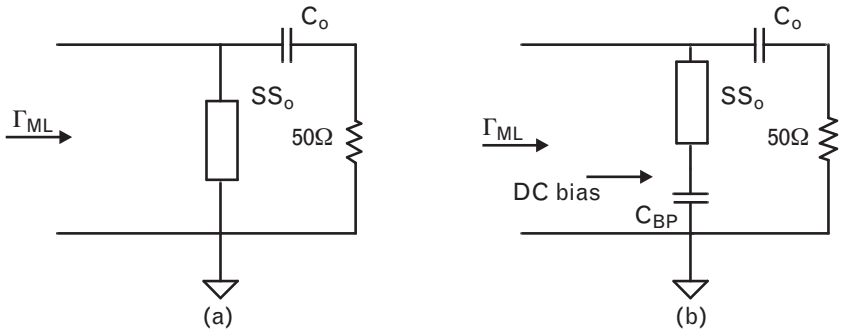


FIGURE 2.4 *(a) The 50-Ω load is transformed to Γ_{ML} at 1.9 GHz with a two-element matching network. (b) By adding an RF short, presented by a resonant capacitor C_{BP}, we can feed dc bias to the collector through the parallel short-circuited matching stub, SS_O. Matching capacitor C_O also serves as a dc block.*

$$C_{\mathrm{OpF}} = \frac{3.183}{f_{\mathrm{GHz}}\, x_{CO}} = \frac{3.183}{1.9(1.92)} = 0.88\ \mathrm{pF}$$

and setting the parallel stub's characteristic impedance arbitrarily to an easily realizable value of 70Ω,

$$\theta_{SSO} = \tan^{-1}\left(\frac{Z_0}{Z_{SSO}b_{SSO}}\right) = \tan^{-1}\left(\frac{50}{70(0.52)}\right) = 53.9°$$

Using standard surface mount components, a 0.88-pF capacitor is too small of a value for tolerance considerations, because the best available tolerances are ±0.1 pF. It could, perhaps, be realized in edge-coupled form on the PC board next to the parallel transmission line stub.

At the input side a similar two-element network is not suitable for dc biasing because Γ_{MS} is located on the *lower-than-50-Ω side* of the Smith chart.[2] An RF matching circuit can be created with two elements only, but if we also want to use the circuit for dc feed-through, we need to add one more short-circuited parallel stub, as shown in Figure 2.5.

Reading the normalized series reactance and parallel susceptances of the input network from Figure 2.3,

$$b_{SSI1} = 1.1, x_{CI} = 0.5, b_{SSI2} = 0.53$$

Computing the input matching network element values,

$$\theta_{SSI1} = \tan^{-1}\left(\frac{Z_0}{Z_{SSI1}b_{SSI1}}\right) = \tan^{-1}\left(\frac{50}{70(1.11)}\right) = 32.8°$$

$$\theta_{SSI2} = \tan^{-1}\left(\frac{Z_0}{Z_{SSI2}b_{SSI2}}\right) = \tan^{-1}\left(\frac{50}{70(0.53)}\right) = 53.4°$$

$$C_{IpF} = \frac{3.183}{F_{GHz}x_{CI}} = \frac{3.183}{1.9(50)} = 3.37\text{ pF}$$

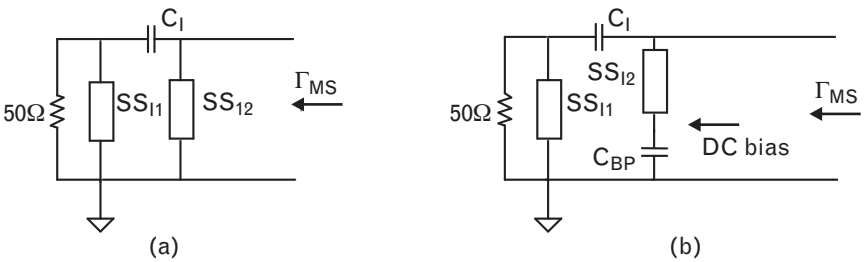


FIGURE 2.5
(a) RF input matching circuit and (b) modified equivalence for dc biasing using an RF short, C_{BP}.

2. A "series capacitor–parallel stub" combination is suitable for transformation to higher impedances.

Now we can put together the complete RF circuit of the 1.9-GHz amplifier, including both matching networks as well as the stabilizing components. The circuit shown in Figure 2.6 is now ready for simulation, using small-signal *S*-parameters to characterize the transistor. All other components are shown with their exact design values and assumed to be lossless initially.

Since the device is unconditionally stable at all frequencies, the transducer gain G_T is equal to the maximum gain, G_{MAX} = 18.2 dB at 1.9 GHz. At that frequency, the amplifier's $|s_{11}|$ and $|s_{22}|$ are zero. Frequency response and port reflection coefficients are displayed in Figure 2.7. Although we designed the impedance matching circuits at 1.9 GHz only, the gain remains flat within 1 dB over a 200-MHz frequency range and both reflection coefficients are less than 0.31 ($VSWR < 2.0$).

For a ±50-MHz range around 1.9 GHz, gain flatness is better than 0.5 dB, and the amplifier's $|s_{11}|$ and $|s_{22}|$ are less than 0.2. Impedance match and gain flatness could be further improved by adding another section to the output matching network. Figure 2.7(a) also shows the effect of component losses.

In addition to component losses, parasitics and layout (including via holes) related effects must also be included in a real design process. At this point the true physical models of the components should replace the initial components. Since that step is customized to the circuit technology and component types, we leave it as an exercise for the reader. In most cases, after the physical models are added, a final optimization is needed to find the final solution.

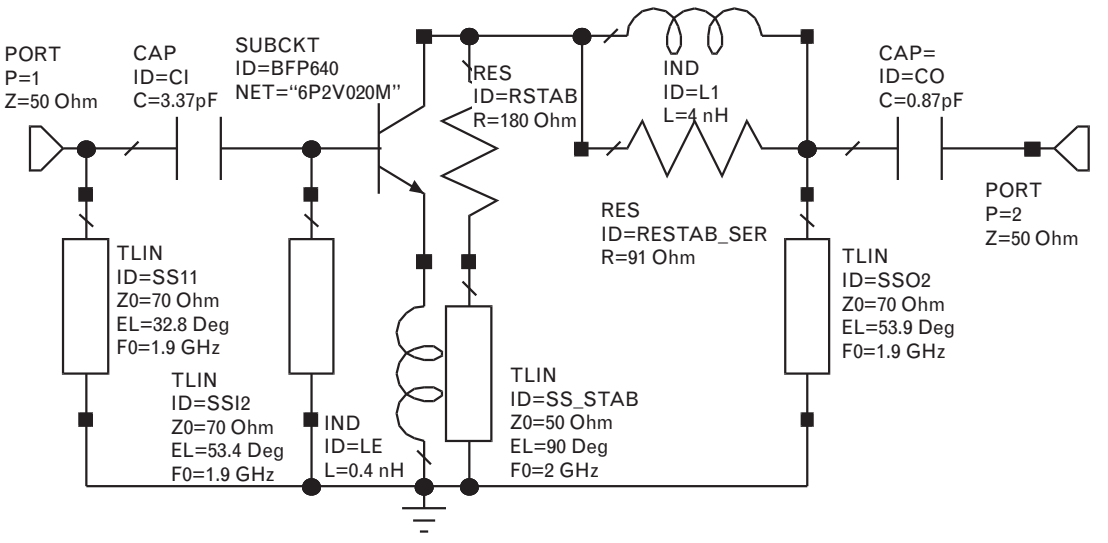


FIGURE 2.6 *RF schematic of the 1.9-GHz amplifier including the device stabilization performed in Section 1.7.*

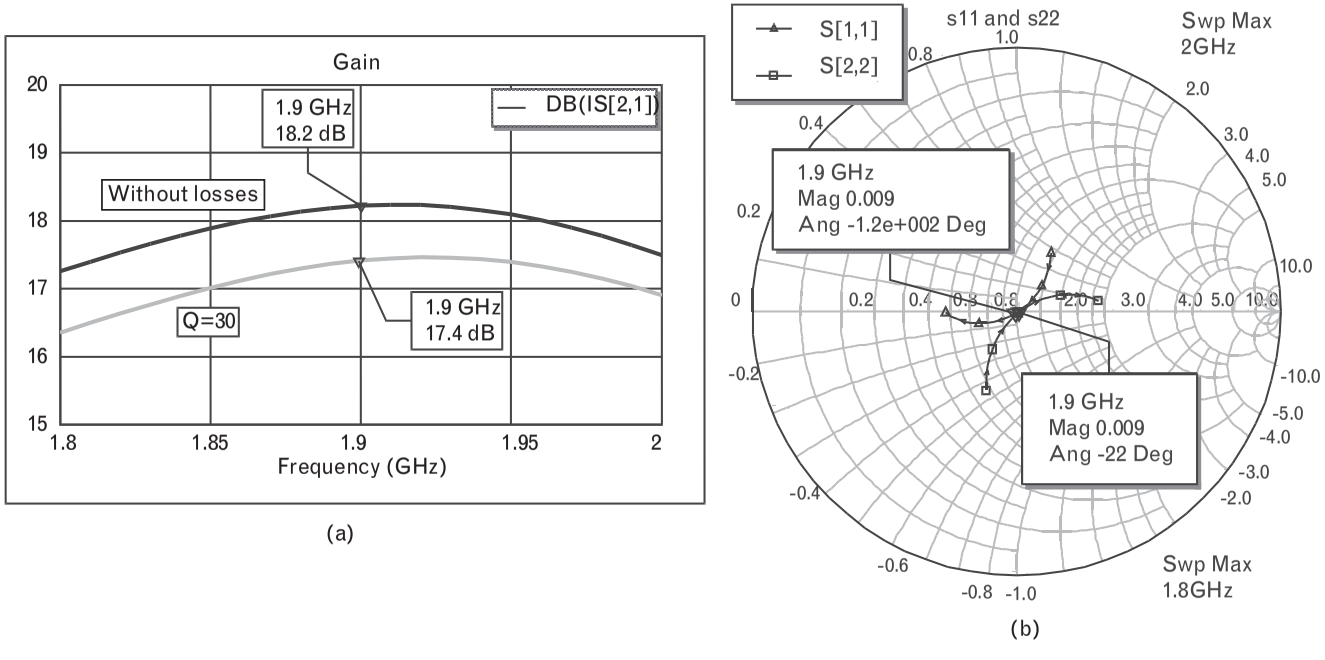


FIGURE 2.7 *(a) Gain and (b) reflection coefficients of the 1.9-GHz amplifier. The lower trace of the gain response shows the reduction caused by changing from lossless components to Q = 30. Losses do not significantly affect impedance match until the component Qs drop to very low levels.*

Component losses generally improve stability, so it is a legitimate question whether we need to fully stabilize the active device before the matching networks are added. Unfortunately, the simultaneous conjugate match formulas only provide workable solutions[3] if the two-port is unconditionally stable; therefore, stabilization is necessary. In the final optimization, we are often able to reduce the attenuation of the stabilizing network(s) to offset part of the component losses of the matching networks.

2.3 Multistage amplifiers

When single-stage amplifiers cannot provide sufficient gain, we can either cascade stages already impedance matched or continuously match the output of one stage to the input of the next stage, until the desired gain is reached. Both techniques have advantages and disadvantages, as described in the next two sections. Although we use 50-Ω external source and load terminations for our examples, the techniques are general and applicable to *any set of arbitrary terminations*.

2.3.1 Cascading impedance-matched stages

When we cascade two perfectly matched, unconditionally stable stages, the overall amplifier will also be stable and impedance matched at both ports, as illustrated for two stages in Figure 2.8. In such a case each individual amplifier's gain in decibels can simply be added to find the overall gain. If the individual amplifiers are not matched, the mismatch between the two stages affects the overall gain.

Since real-life circuits are not perfectly matched, let us examine how the interstage mismatch affects the overall gain. If we know the magnitude and phase of both reflection coefficients that face each other, we can compute the exact amount of mismatch. When we only have the *VSWR*s or

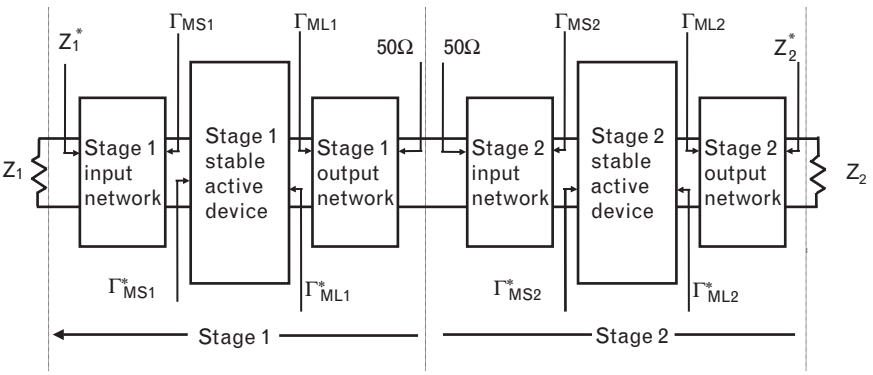


FIGURE 2.8
Cascading two impedance matched two-ports is a quick and simple way to have more gain. The signal level increases from stage to stage, and each succeeding stage must be able to handle the increased signal levels without overloading.

3. For potentially unstable two-ports, the matched terminations have negative real parts.

magnitudes of the reflection coefficients, we can only find the maximum and minimum limits of the interstage mismatch by computing the *mismatch uncertainty*, herein abbreviated as *MU*. Restating *MU* from Volume I, Section 2.14, we have,

$$\textit{Mismatch uncertainty}\ \left(MU_{\text{dB}}\right) = 20\log\left(1 \pm \left|\Gamma_1 \Gamma_2\right|\right) \tag{2.11}$$

where Γ_1 and Γ_2 are the reflection coefficients of the two two-ports being interconnected.

Using (2.11) we find that if the interstage reflection coefficient magnitudes of two cascaded amplifiers are 0.2 each, then *MU* is less than ±0.4 dB. If both magnitudes reach 0.33, *MU* grows to nearly ±1 dB. When several two-ports are cascaded, either active or passive, *MU* becomes very important and system designers need to address the ambiguity factor of the total gain. The problem becomes more significant in wideband communication systems where maintaining low reflection coefficients is very difficult. In such cases we may need to rely on directional passive components, such as isolators, to keep *MU* at reasonable levels.

2.3.2 Cascading amplifiers by direct impedance matching

While cascading individually matched stages offers easy trouble-shooting, a disadvantage lies in more complex interstage networks. Direct impedance matching yields simpler interstage networks but more complicated trouble-shooting, particularly when more than two stages are cascaded. This difficulty is particularly obvious when the active device impedances are far from 50Ω.

Matching one stage directly to the next one is straightforward as long as we remember that the reflection coefficient of any given port and its matched termination are complex conjugates of each other. In a simultaneously conjugate matched amplifier, the load is Γ_{ML}. The output reflection coefficient of the device is simply the complex conjugate of the load,

$$\Gamma_{OUT} = \Gamma_{ML}^{\star} \tag{2.12}$$

The same applies to the input port, where

$$\Gamma_{IN} = \Gamma_{MS}^{\star} \tag{2.13}$$

Table 2.3 summarizes the beginning and ending Smith chart locations for the direct impedance matching of a two-stage amplifier, where both two-ports are unconditionally stable. If the source and load terminations

Table 2.3 Beginning and Final Locations of the Three Matching Networks of a Two-Stage Amplifier with Direct Impedance Matching Technique

Type of Circuit	Start from	End at
Stage 1 input network	50Ω (or any other actual source)	Γ_{MS1}
Interstage matching network	Γ^*_{ML1}	Γ_{MS2}
Stage 2 output network	50Ω (or any other actual load)	Γ_{ML2}

Note: Impedance transformations can be conveniently performed on the Smith chart for lumped or distributed components. If the source and/or load are other than 50Ω, use the actual impedances in column 2.

are 50Ω, the impedance transformations of the input and output networks begin at the center of the Smith chart.

Applying (2.12) and (2.13) to a two-stage cascade, we match the output of the first stage directly to the input of the second stage, as shown in Figure 2.9. In many cases, this approach provides a simpler interstage matching network topology. For example, when cascading two low-impedance devices, such as power transistors, it makes little sense to uptransform the output impedance of one stage to 50Ω and downtransform later to the input of the next stage.

2.3.2.1 Illustrative example: two-stage 1.9-GHz amplifier using direct impedance matching

For simplicity, we use two identical devices under the same bias conditions. (Generally, the second stage is biased at a higher collector current in order to handle the larger signal level.) The input matching network of the first stage and the output network of the second stage are exactly the same as those of the single-stage design, shown in Figures 2.4 and 2.5. Source reflection coefficient of the interstage network is the output reflection

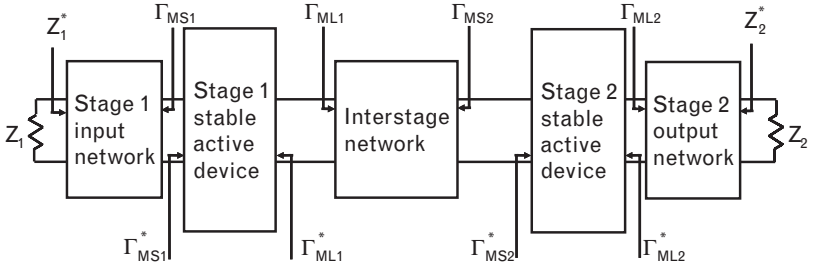


Figure 2.9 *Direct impedance matching offers simpler interstage networks but more complicated trouble-shooting, particularly when more than two stages are cascaded. With lossless matching networks, all interconnected blocks are conjugate matched to each other.*

coefficient of the first stage, $\Gamma_{OUT1} = \Gamma^*_{ML1}$. The interstage network transforms Γ^*_{ML1} to Γ_{MS2}, which is the source needed to conjugate match the input of the second stage.

From Table 2.2 we read Γ_{MS} and Γ_{ML} of the BFP640 at 1.9 GHz to determine the starting and ending points for the interstage network design. Since in our example the two stages are identical, we can use the same parameters for both stages. Therefore,

$$\Gamma_{ML1} = \Gamma_{ML} = 0.65 \angle 13.6°$$

The interstage matching design starts at

$$\Gamma^{\star}_{ML1} = 0.65 \angle -13.6°$$

The ending location is

$$\Gamma_{MS\,2} = \Gamma_{MS} = 0.41 \angle 166°$$

Figure 2.10 shows schematics of the two-stage amplifier with identical stabilized BFP 640 transistors, using direct impedance matching. The three-element PI-network of two short-circuited parallel stubs with a series capacitor between them is also suitable for dc biasing the output of first and input of second stage. Note that the input network of first stage and output network of second stage are the same ones we used in the single stage amplifier of Figure 2.6.

The frequency response of the two-stage amplifier is plotted in Figure 2.11. Since both stages are simultaneously matched, the overall gain at 1.9 GHz is 36.4 dB, exactly twice the gain of the single stage.

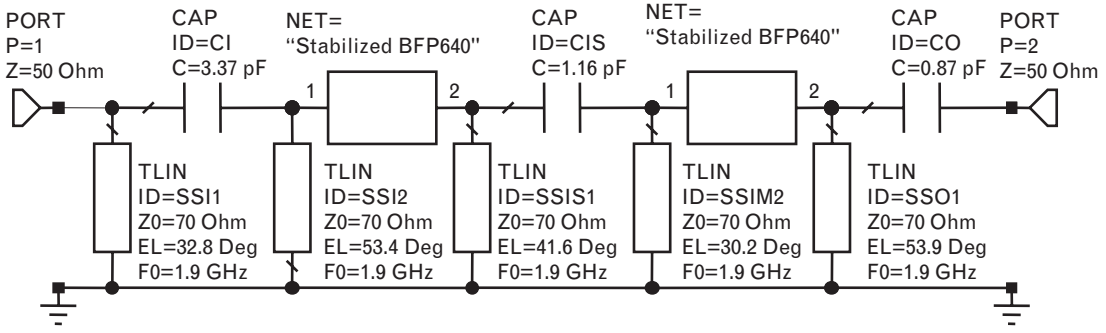


FIGURE 2.10 *Highpass and lowpass interstage network configurations for direct impedance matching between two stabilized BFP 640 transistors.*

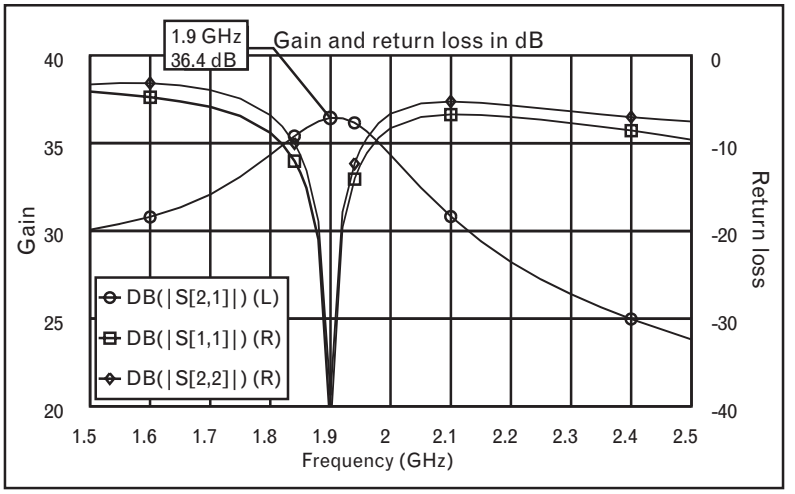


FIGURE 2.11
Gain and return loss plots of the two-stage amplifier show the accuracy of bilateral design. Bandwidth may be improved by increasing the order of the impedance matching networks.

So far, we have covered amplifier design for the maximum gain of unconditionally stable devices using the *transducer gain technique*, derived from the transducer gain expression. Finding a unique set of simultaneously conjugate matched source and load terminations for the active devices gave us excellent input/output match and readily *cascadeable* system blocks.

If minimum noise figure and maximum linear output power could also be reached with simultaneous conjugate matched terminations, the transducer gain approach would take care of all linear amplifier designs. Unfortunately, that is not the case. Minimum noise is not reached when the source impedance is conjugate-matched to the input of an active device [4, 5]. Under real-life physical constraints the absolute output power is less than maximum when a transistor is working into a matched load [6] (for a detailed explanation, see also Chapter 5). Two additional techniques, based on the available power gain [2] and operating power gain [2] expressions, are needed for low-noise and linear power amplifier design.

2.3.3 Output power and impedance match considerations of cascaded amplifiers

For narrowband[4] applications, multistage amplifiers designed for maximum small-signal gain are perfectly matched to each other at band center and reasonably well matched at band edges. In broadband multistage amplifiers, the excessive low-frequency gain is either purposely reflected or dissipated in lossy networks. As a result, some of the stages must generate higher power levels at lower frequencies than at the high end of the passband to overcome the losses. Unless gain equalization is carefully planned

4. Fractional bandwidth is less than 10%.

and distributed through the cascade, we may overload the driver stage(s), and the dynamic range of the amplifier may be severely reduced. We cover nonlinear RF power-level budgeting later in Chapter 5.

Impedance matching represents another problem. When the individual amplifiers are not matched, the interconnecting links may exaggerate the gain ripples of amplifiers. In that case, we can get large ripples and unexpected extreme gain within the passband. To illustrate this problem, we show an extreme example of two identical cascaded bipolar transistor stages with a short segment of 50-Ω transmission line between them (Figure 2.12). Each amplifier has 14.25 ± 0.5 dB gain between 10 and 4,000 MHz in the 50-Ω system.

The block diagram of the amplifier and its broadband frequency response are shown in Figure 2.12. We can clearly see that the overall gain is not twice the decibel gain of the individual cascaded amplifiers.

The amplifier stages of Figure 2.12 consist of Infineon BFP 520 transistors, biased at 2V, 20 mA, with 16-Ω resistive series feedback applied. The feedback maintains flat broadband gain and nearly unity magnitude input/output reflection coefficients. At some frequency, the two high-impedance amplifiers match into each other and we get *more than* 54 dB total gain. At another frequency where the two stages see the largest interstage mismatch, the overall gain drops to *under* 23 dB. If we change the length of the 50-Ω interstage transmission line, the peaks and dips of the gain response also move.

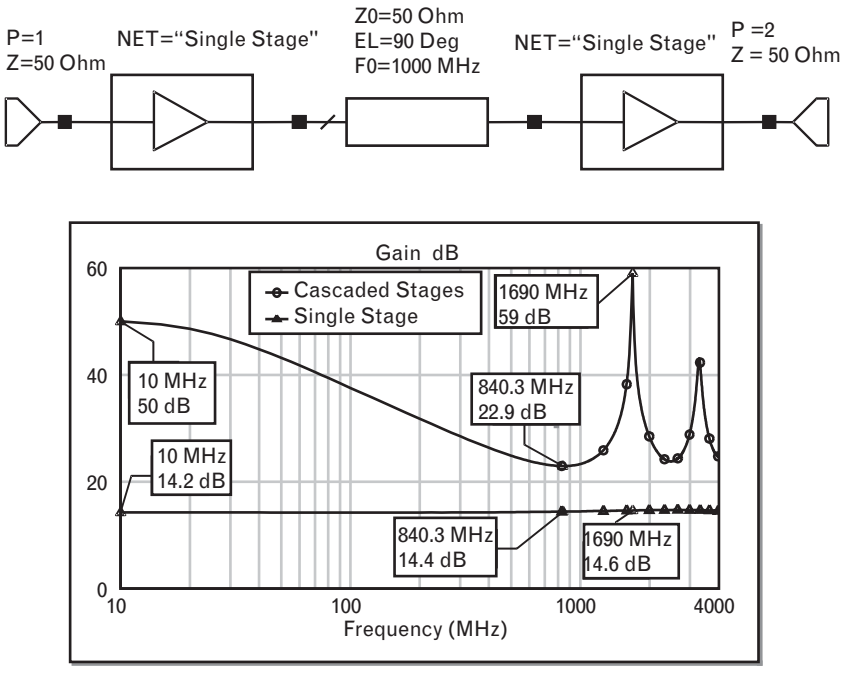


FIGURE 2.12
Do we get twice as much gain by cascading two identical amplifiers? Not in this case where the input and output impedances of the stages are far from 50Ω.

2.4 Operating gain design for maximum linear output power

The *operating gain* design approach starts from the desired load impedance and then matches the resultant input impedance (see Figure 2.13). The operating gain technique is recommended for linear-power amplifiers, where the load is the more important of the two terminations. Amplifiers designed this way have *only one matched port* (input port). Since the output port is not matched we do not get the maximum small-signal gain, but that is the price we pay to get *maximum absolute output power.*

Operating power gain, derived from the transducer gain equation, is a function of the *S*-parameters and the load reflection coefficient of a two-port. For a given set of two-port *S*-parameters, the operating gain is a function of Γ_L only [2], that is,

$$G_p = \frac{\text{Power delivered to the load}}{\text{Power applied to the imput of the two-port}} = \frac{|s_{21}|^2\left(1-|\Gamma_L|^2\right)}{\left(1-\left|\dfrac{s_{11}-(\Delta)\Gamma_L}{1-s_{22}\Gamma_L}\right|^2\right)|1-s_{22}\Gamma_L|^2} \tag{2.14}$$

where the symbol Δ is the determinant of the two-port *S*-matrix,

$$\Delta = s_{11}s_{22} - s_{12}s_{21} \tag{2.15}$$

In contrast with the G_T definition, the source reflection coefficient is not part of the operating power gain expression. G_p definition involves the *applied input power* of the two-port, which is independent of Γ_s. For example, if an amplifier has a power gain of $G_p = 10$ and the input power is 10 mW, the output power is 100 mW. If conjugately matching the input

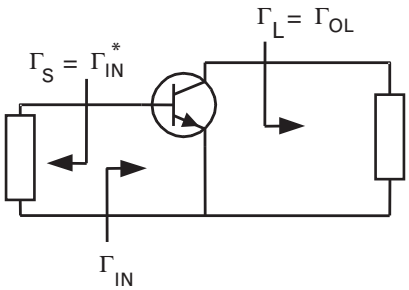


FIGURE 2.13 *Source and load terminations of the operating power gain design technique. The source is always conjugate-matched to the input impedance of the two-port. The load is selected for special considerations, such as maximum output power,* Γ_{OL}.

increases the applied input power to 20 mW, the output power is 200 mW. Obviously, when designing for a given operating power gain, it helps to have the input port conjugately matched in order to obtain the largest input and output power levels. Furthermore, when the input port is conjugately matched there is no mismatch loss at the input. Therefore, the input VSWR = 1. If the *S*-parameters of a two-port are known, we can find G_p for any specified Γ_L.

Solving (2.14) for Γ_L results in the equations for a family of *operating constant-gain circles* for various values of G_p. A particular gain circle represents the locus of load terminations leading to that specific power gain of G_p.

If the two-port is unconditionally stable, the operating gain circles for various gain levels are completely inside a Smith chart. For potentially unstable two-ports, a part of each circle lies outside of the same chart. The arc of the unstable region on the Smith chart is marked by the intersection of the load stability circle with the operating gain circles as shown in Figure 2.14.

2.4.1 Operating gain design outline

The single-frequency linear amplifier design procedure with the operating gain technique (covered in Chapter 5) is as follows:

- The first requirement is to determine the load, Γ_L, for the device. Once the desired load is defined, the corresponding small-signal gain is computed from (2.14).
- If the gain is not sufficient, we can plot constant operating gain circles and constant-power contours [7, 8] to investigate possible trade-offs

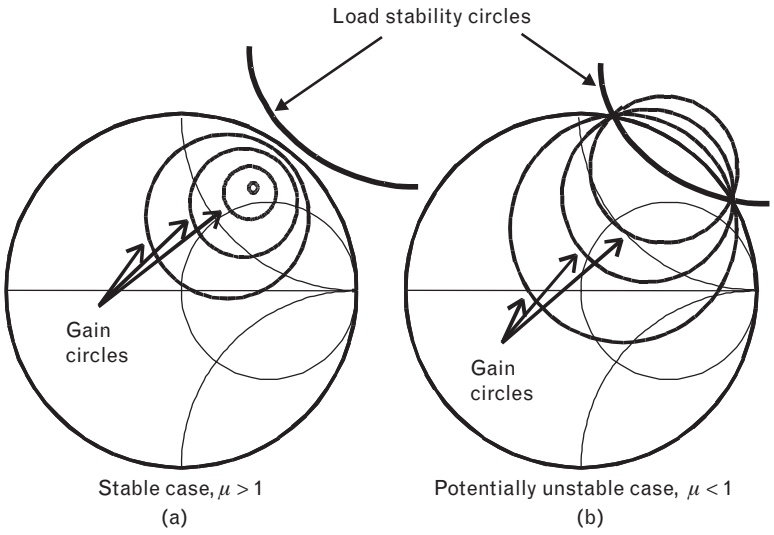


FIGURE 2.14
Bilateral operating gain circles represent load terminations that provide equal gain when the input of the two-port is conjugate-matched. (a) For stable devices the gain and load stability circles do not intersect each other. (b) Potential instability exists when parts of gain circles are intersected by the load stability circle.

between gain and output power. Constant output power contours may be approximated [6] from dc bias parameters. Exact plots require a nonlinear device model, load-pull information, and nonlinear circuit simulation, which are covered later in Chapter 5.

- Design an output circuit that transforms the existing load termination to this new Γ_L.
- Connect the new Γ_L to the device and calculate the resulting input reflection coefficient,

$$\Gamma_{IN} = s_{11} + \frac{s_{21}s_{12}\Gamma_L}{1 - s_{22}\Gamma_L} \tag{2.16}$$

- Since the operating gain approach is based upon a conjugate-matched input port, create an input circuit to transform the system termination to the required conjugate matched source,

$$\Gamma_S = \Gamma_{IN}^{*} \tag{2.17}$$

- Place the device between the two new terminations, Γ_S and Γ_L. Using lossless matching elements, the amplifier now has the exact predicted gain and a matched input port. The output port is *not matched* since the amplifier's gain is less than G_{MAX}.

A linear power amplifier block diagram, describing the functions of the input and output networks, is shown in Figure 2.15. The output network provides optimum loading for absolute output power. The function of the input network is to match the input port for maximum input power. In broadband applications the input network may also be used as a gain-equalizer.

This technique works well, regardless of the RF stability condition of the device. With an unconditionally stable active two-port, the highest achievable gain is G_{MAX}, in which case the source and load terminations

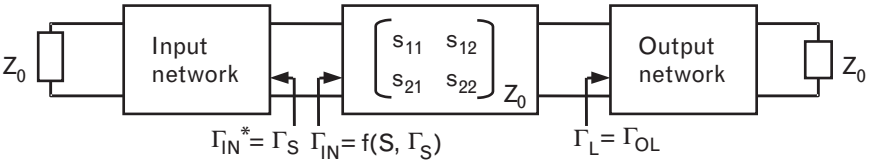


FIGURE 2.15 *Generalized block diagram of a two-port connected to arbitrary source and load terminations. The two-port is characterized by its small-signal S-parameters. The input port is conjugate matched to Z_0, but the actual load is transformed to the optimum load needed for a specific goal, such as maximum absolute output power.*

revert to the unique set of values: $\Gamma_S = \Gamma_{MS}$ and $\Gamma_L = \Gamma_{ML}$ (i.e., the simultaneous conjugate matched values). For all other gains below G_{MAX}, an infinite number of source and load combinations always provide matched inputs and mismatched outputs.

2.4.2 G_P versus P_{OUT} trade-offs

Just as we face trade-offs between noise performance and gain at the input of low-noise circuits, similar trade-offs between the small-signal gain and maximum output power exist at the output of an active two-port [6]. When the *constant-output-power contours* and *operating constant-gain circles* are superimposed on the Smith chart, we can choose the load either for maximum gain or maximum output power, or for a compromise between these two extremes. A plot of constant gain and constant power input will be shown in Section 2.4.4. The input port is matched in all cases. Other considerations, such as harmonic and intermodulation distortion, and power-added efficiency may also be brought into the decision-making process.

2.4.3 Stability considerations

RF stabilization of power amplifiers is still somewhat of an unsettled issue among power amplifier designers [6]. Since virtually all of the modern RF power devices are potentially unstable, simultaneous conjugate match cannot be achieved with realizable passive circuits. In power amplifiers, however, the output port is not matched, but instead it is terminated for maximum output power. Still, when the input port is conjugate-matched to the source, the output reflection coefficient magnitude in some cases may be greater than unity. Under such conditions, depending on the actual load impedance seen by the device outside of the passband, oscillation may take place.

Even though the procedure outlined in Section 2.4.1 may show no problem through the first three steps, we may find that the matched source reflection coefficient lies very close to the unstable source region. Choosing such a source leads to an output reflection coefficient magnitude just under unity, which is still an undesirable situation. A safer approach may be to stabilize the device first and sacrifice gain. Since absolute power output is very important, we must perform stabilization at the input port.

If an amplifier is designed for a known system, the broadband source/load terminations facing the active device may either be measured or simulated. Then, RF stability needs to be assured only for the existing terminations. Accordingly, the amplifier may be designed with a potentially unstable device. Since the operating gain technique also works with potentially unstable devices, we may proceed without stabilization.

If the amplifier is aimed at the *original equipment market* (OEM), the actual broadband source/load terminations are not known. In such a case, it is a very important to stabilize the device at all frequencies to prevent possible oscillation. Component buyers generally specify unconditional stability, not just at all frequencies but also at various input power levels. *S*-parameter–based stability analysis is not adequate for high power applications and requires true nonlinear simulation using Nyquist stability criterion discussed here and in Chapter 4.

2.4.4 Illustrative example: operating gain design for maximum linear power output

Find the required terminations for maximum linear output power at 1.95 GHz, using the NEC6500379A GaAs power MESFET, operating at 3V, 800-mA dc bias, into 50-Ω RF source and load terminations.

Solution: Small-signal *S*-parameters of the active device are listed in Table 2.4. The device is potentially unstable through a wide range of frequencies. Since we do not want to add any loss to the output side of the device, we need to find out how it can be stabilized at the input port.

Plotting the source stability circles (Figure 2.16) reveals that we need to avoid low-impedance source terminations. Unfortunately, a low-impedance device, such as a power transistor, generally requires a low-impedance source for maximum gain since its input impedance is quite low. We may need to sacrifice some gain to have unconditional stability.

To illustrate the operating gain design procedure, we determine the required terminations of a 1.9- to 2.0-GHz linear power amplifier using this MESFET. The optimum load impedance for maximum linear power output at 1.95 GHz is $Z_{OL} = (3.2 - j5)\Omega$ (a more detailed discussion of determining the optimum load and large-signal stability follows in Chapter 5). If possible, we also want to stabilize the device because it is potentially unstable through a wide frequency range, including our target frequency.

Table 2.4 *S*-parameters of the NEC NE6500379A (Biased at 3V, 800 mA) at 1.95 GHz, without and with Stabilization

	s_{11} Mag	Ang	s_{21} Mag	Ang	s_{12} Mag	Ang	s_{22} Mag	Ang	μ	MSG/G_{MAX} dB	Z_{OL}-Ω Rg	Im
Before stabilization	0.96	166	0.596	69	0.28	−11	0.91	169	0.84	13.2	3.2	$-j5$
After stabilization	0.87	167	0.567	70	0.027	−10	0.91	168	1.02	10.0	3.2	$-j5$

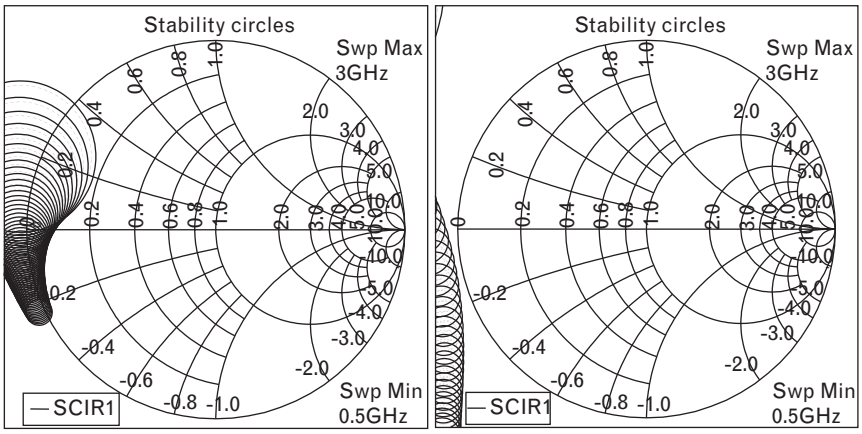


FIGURE 2.16
Source stability circles of the NE6500379A through the 0.5- to 3.0-GHz frequency range before and after stabilization. Unstable regions are inside of the stability circles.

The lossy network combination shown in Figure 2.17 stabilizes the device with only a small amount of gain sacrifice at 1.95 GHz. The 20-Ω series resistor provides broadband stability at a very large reduction of gain. Adding the parallel branch that resonates at 1.95 GHz helps to recover some of the gain in the passband but still maintains stability at other frequencies.

Our example is based on ideal passive RLC elements for stabilization. A physical circuit's performance depends on the specific components, and the modeling task is left to the reader as an application exercise.

RF stabilization always reduces small-signal gain. In our example, as shown in Table 2.4, before stabilization the maximum stable gain, *MSG*, is 13.2 dB. After stabilization, the highest gain is 10 dB, which means we need to give up over 3-dB gain at 1.95 GHz for stability. The question is whether it is worth it to sacrifice gain to have RF stability and a better output match. This is a choice the designer must make.

When the stabilized device is *tuned for maximum output power*, the gain drops further to 9.1 dB, as shown in Figure 2.18. Since the margin of RF stability is quite small (i.e., $\mu = 1.02$), after the input port is matched, the

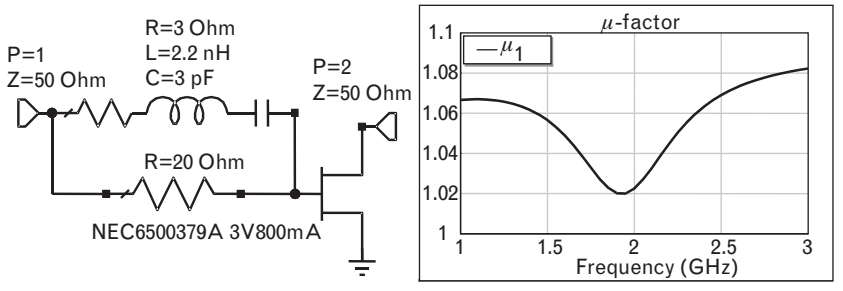


FIGURE 2.17 *Frequency selective network used to stabilize the NE6500379A FET. The gain reduction caused by the series 20-Ω resistor is minimized at 1.95 GHz with the help of the series RLC network. Broadband stability is assured with the 20-Ω resistor outside the passband. Simulation is based on ideal components.*

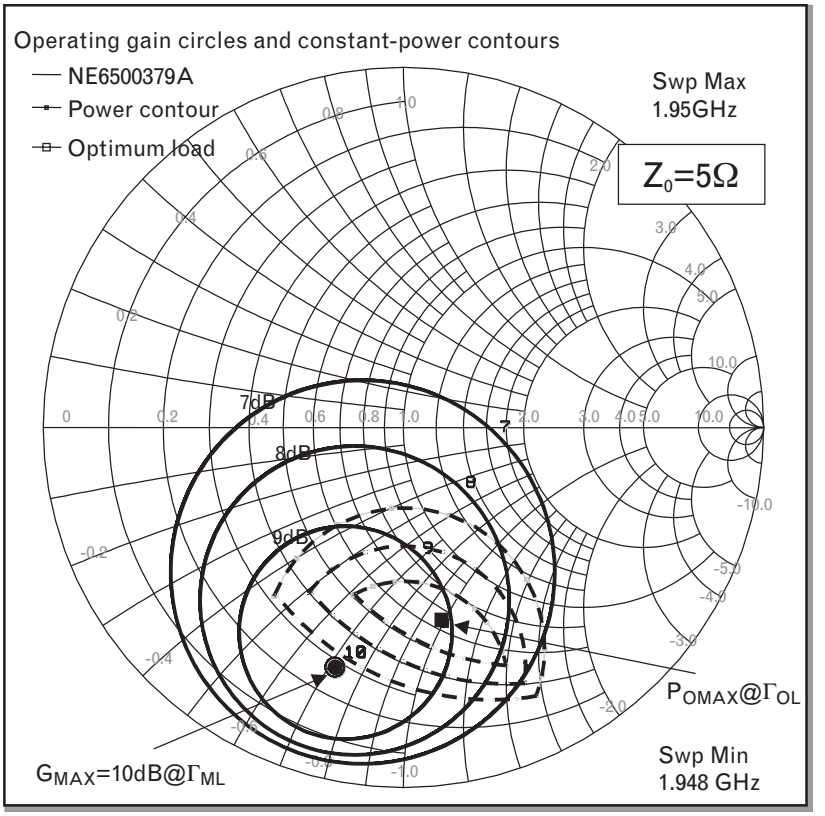


FIGURE 2.18
Operating gain circles and approximated constant output power contours in 1-dB steps show the trade-offs between small-signal gain and power output. At maximum power (Γ_{OL}) the small-signal gain is 9.1 dB. When the output is matched for maximum small-signal gain (Γ_{ML}), the output power is more than 3 dB down from P_{OMAX}. Note: Smith chart is normalized to 5Ω for good resolution.

output reflection coefficient magnitude of the stabilized device may still be high—but no longer exceeds unity, as it would without stabilization.

Using the stabilized parameters of Table 2.4, we now illustrate the *operating gain* design with the procedure outlined in Section 2.4.1.

- Convert the optimum load for maximum output power to reflection coefficient, Γ_{OL}

$$\Gamma_L = \Gamma_{OL} = \frac{Z_{OL} - Z_0}{Z_{OL} + Z_0} = \frac{(3.2 - j5) - 50}{(3.2 - j5) + 50} = 0.88 \angle -168.3°$$

- Compute the new input reflection coefficient of the stabilized device with $\Gamma_L = \Gamma_{OL}$, using (2.16):

$$\Gamma_{IN} = s_{11} + \frac{s_{21}s_{12}\Gamma_{OL}}{1 - s_{22}\Gamma_{OL}} = 0.88 \angle 171°$$

- Now, set the source reflection coefficient at the complex conjugate of Γ_{IN}, as given in (2.17):

$$\Gamma_S = \Gamma_{IN}^{\star} = 0.88 \angle -171°$$

- Performing small-signal simulation of the two-port shown in with the computed source and load at 1.95 GHz gives us the *S*-parameters of the power amplifier (see Table 2.5).

2.4.5 Output match considerations

Using the operating gain design approach produces amplifiers with well-matched input and mismatched output. For a stable two-port, the quality of output match depends on how far Γ_{OL} is from Γ_{ML}. For a potentially unstable device, the situation is more complicated, and the output reflection coefficient of the amplifier may be outside the Smith chart at some frequency. In such a case, we may need to place a directional element[5] at the output to "cover up" the poor output match. These elements add more loss to the output side, which is not desirable in power amplifiers. Balanced amplifiers, covered in Section 2.6.2, use directional couplers to hide the poor input match caused by low-noise design. The same approach may also be used for power amplifiers if the output match is a serious problem. The losses of the directional couplers always reduce gain, but the absolute maximum output power increases, due to the added transistor, as we will see in Section 2.6.4.

In summary, the operating gain procedure is directly applicable to linear power amplifier design since it allows us to choose the load for maximum absolute output power. The technique works regardless of the RF stability condition of the active device. If we are assured that oscillation will not take place, unconditional stability may be sacrificed to satisfy output power, linearity, and gain requirements. A more detailed discussion of those topics is given in Chapter 5.

Table 2.5 *S*-Parameters of the Amplifier Show Perfect Input Match and a Mismatched Output Port

Frequency (GHz)	s_{11} Mag	Ang	s_{21} Mag	Ang	s_{12} Mag	Ang	s_{22} Mag	Ang		
1.95	0.00	55	2.86	79	0.13	–1	0.39	125	1.02	9.1

Note: Since the device is terminated for maximum absolute output power, the small-signal gain is only 9.1 dB.

5. Such as an isolator or directional coupler.

2.5 Noise in RF circuits

Before proceeding with the available power gain definitions, let us look at the sources that create noise in RF circuits and define various noise-related circuit and device parameters. Although we already cover some of these definitions in Volume I, Chapter 3, a short review is still useful for those who do not have immediate access to that book.

2.5.1 Review of noise sources in RF systems

At the input of an RF receiver the signal levels may be extremely low, and we need to minimize the internal noise generated by the system. The three main causes of electrical noise listed in Volume I, Sections 3.3 and 3.4, are:

- *Thermal*, or Johnson noise, caused by the thermal agitation of free electrons in conductors;
- *Shot*, or Shottky noise, caused by the random fluctuation of current flow in semiconductors;
- *Flicker*, or $1/f$ noise, caused by fluctuation in the conductivity of the medium.

The first two noise types are broadband[6] but the $1/f$ noise is only a concern when the passband reaches the low megahertz region (unless the device is being used for upconversion or downconversion, as in a mixer or oscillator). Noise is a random phenomenon, and at RF we prefer to deal with noise power (instead of noise voltage or noise current) that may be combined from different sources.

In addition to internally generated noise, there are also external noise sources [9], such as atmospheric, galactic, solar, ground, and man-made noise. Since those are not circuit related and may be out of our control, we do not cover them in this text.

In Volume I, Section 3.3, we also define the *noise factor* of a two-port as

$$F = \frac{\text{Actual noise power at the output of the two-port}}{\text{Expected noise power at the output of the ideal (noiseless) two-port}} = \frac{\text{Signal-to-noise ratio at the input}}{\text{Signal-to-noise ratio at the output}} \tag{2.18}$$

6. Also referred to as white noise.

The *noise figure* in decibels is

$$NF = 10\log(F) \tag{2.19}$$

Another form of evaluating noise performance is the *noise measure*, M,

$$M = \frac{F-1}{1-\frac{1}{G_A}} \tag{2.20}$$

The noise measure takes into account of the gain of the two-port. If G_A, the available gain of the two-port, is at least 15 to 18 dB, then $M \approx F$.

One more related figure of merit is the *noise temperature*, often used by antenna designers:

$$T = T_0(F-1) \tag{2.21}$$

where T_0 is 290K.

A noiseless two-port has unity noise factor and 0-dB noise figure. Cascading two or more noisy two-ports, the overall noise factor of N-stages, denoted by $F_1, F_2, \ldots, F_N$, is given by [10]

$$F = F_1 + \frac{F_2 - 1}{G_{A1}} + \ldots \frac{F_N - 1}{G_{A1}G_{A2}\ldots G_{A(N-1)}} \tag{2.22}$$

Equation (2.22) shows that, in addition to the noise factors themselves, the gains of the first and second stages may also be important in keeping the overall noise factor low.

We should point out that in industrial practice the component's noise and gain are generally specified in decibel values. Most of the RF noise formulas in textbooks use noise and gain in power ratios instead of decibels. Before we apply real-life specifications to formulas like (2.22), the decibel values must be converted to power ratios. For a deeper discussion of noise-related topics, we refer the reader to [11].

The formula may be applied by recalling from Volume I, Section 2.2, the conversion from decibels to power ratio:

$$\text{Power ratio} = 10^{\frac{\text{dB value}}{10}} \tag{2.23}$$

Input circuit losses must be minimized because they directly contribute to the noise of the system [12]. To illustrate this point, let us look at a

two-stage LNA combination with a lossy 50-Ω cable connected to its input (Figure 2.19). For simplicity, assume that the amplifiers are perfectly matched to 50Ω at all ports.

2.5.1.1 Illustrative exercise: cascade noise figure calculations

Let us compute the noise factors and gain of the three cascaded blocks of Figure 2.19, using (2.23):

$$F_1 = 10^{\frac{NF_1}{10}} = 2 \quad F_2 = 10^{\frac{NF_2}{10}} = 1.41 \quad F_3 = 10^{\frac{NF_3}{10}} = 2.51$$

$$G_{A1} = 10^{\frac{\text{Gain}_1}{10}} = 0.5 \quad G_{A2} = 10^{\frac{\text{Gain}_2}{10}} = 100$$

Then, the overall *noise factor*, *F*, from (2.22) is

$$F = F_1 + \frac{F_2 - 1}{G_{A1}} + \frac{F_3 - 1}{G_{A1}G_{A2}}$$

$$= 2 + \frac{1.41 - 1}{0.5} + \frac{2.51 - 1}{50} = 2 + 0.82 + 0.03 = 2.85$$

Converting to *noise figure* with (2.19)

$$NF = 10 \log (2.85) = 4.55 \text{ dB}$$

Of the 3.05-dB noise figure increase, 3 dB is contributed by the lossy cable. If the circuit is rearranged by moving the LNA to the front as shown in Figure 2.20, the new noise figure becomes significantly lower.

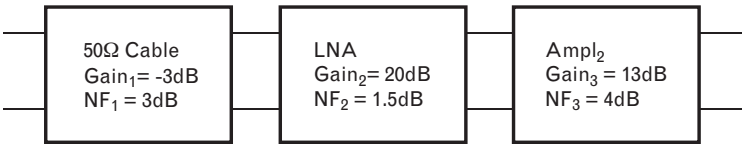


FIGURE 2.19 *Two-stage LNA is preceded by a 50-Ω cable that has 3-dB total attenuation. The cable loss directly increases the overall noise figure. The overall noise figure increase is 1.55 dB.*

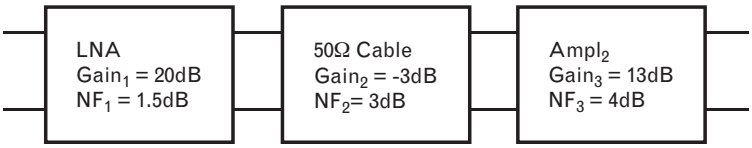


FIGURE 2.20 *Moving the LNA to the input of the three-element cascade keeps the overall noise figure at a low level. The overall noise figure increase is 0.11 dB.*

$$F_1 = 10^{\frac{NF_1}{10}} = 1.41 \quad F_2 = 10^{\frac{NF_2}{10}} = 2 \quad F_3 = 10^{\frac{NF_3}{10}} = 2.51$$

$$F = F_1 + \frac{F_2 - 1}{G_{A1}} + \frac{F_3 - 1}{G_{A1}G_{A2}}$$
$$= 1.41 + \frac{2-1}{100} + \frac{2.51-1}{50} = 1.41 + 0.01 + 0.03 = 1.45$$

Therefore, the new noise figure is

$$NF = 10 \log (1.45) = 1.61 \text{ dB}$$

Now the noise figure of the LNA is only increased by 0.11 dB—quite an improvement from the previous interconnection scheme. The small increase is due to the high gain of the LNA. If the gain of the LNA in the second case is decreased to 14 dB, the overall noise figure increases to 1.96 dB. Using only 10-dB gain for the LNA gives us overall $NF = 2.58$ dB. Having at least 15- to 18-dB gain for the LNA generally helps to overcome the second-stage noise contribution. Of course, there are exceptions, and here is an example that happened in the 1960s.

A two-stage broadband LNA was developed to convert the 50-Ω input impedance of an early-day spectrum analyzer to high impedance and low- capacitance with an active probe. As the proud designer, I announced in a project review that the new probe will allow engineers to see low-level signals, since the broadband probe's noise figure was around 4 dB with 17-dB gain. My morale was considerably lowered after learning the noise figure of the spectrum analyzer was 27 dB, increasing the probe-equipped system's overall noise figure to 11 dB. After redesigning the probe by adding a third amplifier stage, its noise figure increased to 4.1 dB and, more importantly, the gain to 25 dB. The higher front-end gain of the probe helped to lower the overall system noise figure to 6.2 dB.

Generally, after two stages of amplification, the signal and noise levels are high enough and further stages do not affect the overall noise performance. Exceptions do exist, however, as stated above where the spectrum analyzer's high noise figure overwhelmed the preamplifier. Having more gain at the front helps, but too much gain in the LNA(s) may overload the next system block when a strong signal appears at the output of the LNA.

Here are some important points to remember:

- Dissipative component losses at the input always degrade noise performance.
- Gain of the first two stages, particularly that of first stage, also affects the overall noise figure.

- Always check the dynamic range[7] of the LNA to prevent distortions in the front-end of the receiver.

2.5.2 Two-port noise parameter definitions

Above absolute zero temperature, even with the input signal of the circuit shown in Figure 2.21 set to zero ($v_S = 0$), a physical two-port always generates some measurable noise. At a given frequency and temperature, the total noise contribution of such a two-port may be represented by a correlated pair of noise-voltage and noise-current generators [2].

The total noise power at the output is highly dependent on the source impedance of the two-port. A unique *optimum source impedance* (Z_{OPT}) exists, that leads to the best noise performance. It can be found by placing a variable source at the input and measuring the noise at the output, as shown in the simplified schematic of Figure 2.22. At low frequencies Z_{OPT} is real, but it becomes a complex impedance above 50 to 100 MHz for most active devices.

The noise factor measured with the source set to Z_{OPT} is called F_{MIN}[8] (or NF_{MIN} when converted to decibels). Z_{OPT} is frequently converted to *optimum noise reflection coefficient*, called Γ_{OPT}. A third noise parameter called *equivalent noise resistance*, r_N, is a sensitivity factor; it shows how fast *NF* increases as the source termination changes from $\Gamma_{OPT.}$ At a specific set of operating conditions, the three noise parameters (F_{MIN}, $\Gamma_{OPT,}$ and r_N) can fully characterize the noise performance of a given two-port. Changing the temperature, frequency, and the dc bias conditions of the active device also change the noise parameters. The noise figure of a two-port network is given by [13]

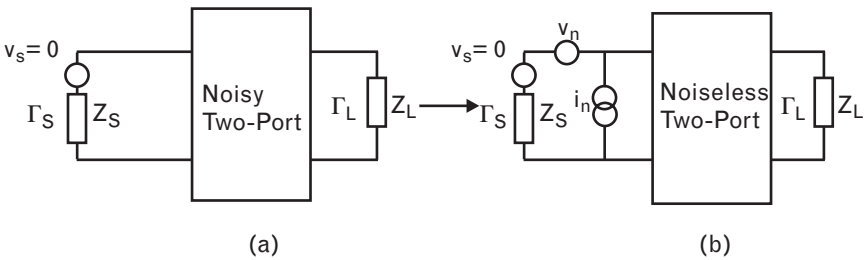


FIGURE 2.21 *(a) Representation of a physical (noisy) two-port, and (b) equivalent circuit for noise considerations by taking the internal noise sources out of the two-port and noise sources to a noiseless two-port.*

7. Difference between lowest detectable and highest allowable signals (see Volume I, Section 2.6).

8. Technically speaking, *F* stands for noise factor, but on transistor datasheets F_{MIN} refers to decibel value instead of power ratio.

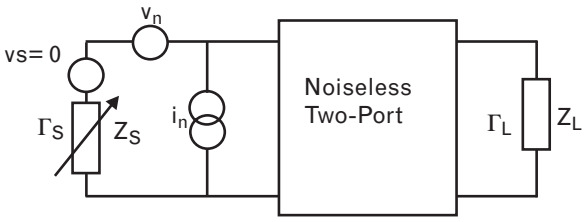


FIGURE 2.22 *Varying the source termination affects the individual noise contribution of the equivalent two noise generators placed the input. The source impedance leading to the lowest noise power at load is called the optimum noise source impedance, $Z_S = Z_{OPT}$, and when converted to a reflection coefficient it is called Γ_{OPT}.*

$$F = F_{MIN} + \frac{4r_N \left|\Gamma_S - \Gamma_{OPT}\right|^2}{\left(1 - \left|\Gamma_S\right|^2\right)\left|1 + \Gamma_{OPT}\right|^2} \tag{2.24}$$

Γ_{OPT} is *not* a function of the small-signal *S*-parameters. For optimum noise performance, the source reflection coefficient is not matched to the input—it is transformed to Γ_{OPT}. Therefore,

$$\Gamma_S = \Gamma_{OPT} \neq \Gamma_{MS}$$

Since the input port is *not conjugate-matched*, in low-noise amplifiers we do not get the maximum gain of the two-port.

Reactively matching or mismatching the output port *does not have any effect* on the signal-to-noise ratio and noise figure. Output matching, of course, provides more gain, which helps to reduce the noise contribution of the next stage.

In multistage low-noise amplifiers, the goal is to minimize the overall noise performance. Ideally, all stages should see their optimum noise sources at their inputs, but that may not lead to minimum overall noise. Circuit optimization is very helpful here to target minimum noise, flat gain response, and good output match simultaneously.

2.6 Available gain design technique

The available gain design technique, a virtual mirror-image form of the operating gain approach, begins at the source termination and requires a conjugate-matched load at the output side, as shown on Figure 2.23. The available gain technique is bilateral since it includes the input-output interaction caused by s_{12}.

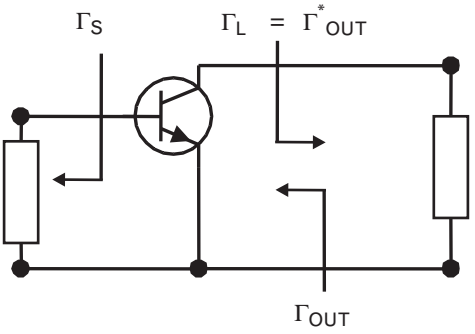


FIGURE 2.23
The available gain approach begins by selecting the source termination first, generally for low noise considerations, and matching the new output impedance to the load.

The mathematical form of available gain, G_A, resembles (2.14) by interchanging s_{11} and s_{22}, and substituting Γ_S for Γ_L.

$$G_A = \frac{\text{Power available from the two-port}}{\text{Power available from the source}} = \frac{|s_{21}|^2\left(1-|\Gamma_S|^2\right)}{\left(1-\left|\frac{s_{22}-(\Delta)\Gamma_S}{1-s_{11}\Gamma_S}\right|^2\right)|1-s_{11}\Gamma_S|^2} \tag{2.25}$$

Solving (2.25) for Γ_S, we get another set of circles, called *constant available gain circles*, for various values of G_A. For a given set of two-port *S*-parameters, the circles are defined by the specified value of available gain. Each circle represents the loci of all source impedances that provide a constant available gain.

Just as with the operating gain circles, the available gain circles are also located completely inside the Smith chart for an unconditionally stable two-port. If the two-port is potentially unstable, the circles are partially outside the Smith chart (see Figure 2.24). Although designing amplifiers with potentially unstable devices may not be a good practice, the available gain technique is usable regardless of stability.

Superimposing available gain circles and constant noise circles on the same Smith chart enables us to see the trade-offs between the noise performance and gain of a two-port. Since the available gain design is a bilateral technique, the load should be conjugate-matched to the output port in order to deliver maximum output power to the load.

2.6.1 Available gain design outline

The single-frequency linear amplifier design procedure with the available gain technique is as follows:

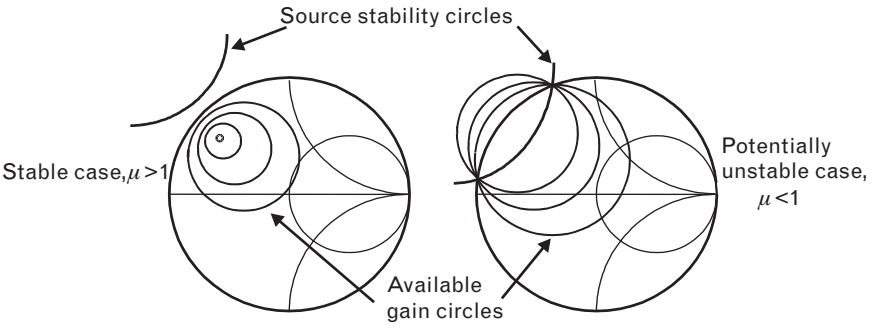


FIGURE 2.24 *Available gain circles are similar to the operating gain circles, except the former show gain variation versus source impedance while the latter refers to load terminations. Left-side plots are for an unconditionally stable two-port. Circles on the right-side plots indicate that the two-port is potentially unstable.*

- As we mentioned, the design process begins at the input side of the two-port. Once the source termination is selected, the corresponding small-signal gain is computed from (2.25). If the gain is not sufficient, we can plot constant available gain and constant-noise circles to investigate possible trade-offs between gain and noise. Once the desired source is defined, we design an input circuit that *transforms* the existing source termination to this new Γ_S.
- Connect the new Γ_S to the device and calculate the resulting output reflection coefficient,

$$\Gamma_{OUT} = s_{22} + \frac{s_{21}s_{12}\Gamma_S}{1 - s_{11}\Gamma_S} \tag{2.26}$$

- Since the available gain approach is based upon a conjugate-matched output port, create an output circuit to transform the system termination to the required conjugate-matched source.
- The final step is to transform the actual system load termination to the complex conjugate of this new Γ_{OUT}:

$$\Gamma_L = \Gamma_{OUT}^{*} \tag{2.27}$$

- Place the device between the two new terminations, Γ_S and Γ_L. Using lossless matching elements, the amplifier now has the exact predicted gain and a matched output port. The input port is not matched since the amplifier's gain is less than G_{MAX}.

Amplifiers designed with the available gain technique have perfectly matched output and a mismatched input port. The amount of mismatch

loss at the input determines the magnitude of the input reflection coefficient. For example, if we have to sacrifice 1-dB gain at the input port for the best noise figure, the 1-dB mismatch loss converts to a 0.45 input reflection coefficient magnitude. A 2-dB mismatch loss leads to an input reflection coefficient magnitude of 0.6, which is a very poor input match. For that reason, we need to look for special techniques to maintain reasonable input match for low-noise amplifiers.

2.6.2 Low-noise amplifier design considerations

For most small-signal RF/MW transistors, the gain at minimum noise conditions is about 3 to 6 dB below the maximum gain of the device, for two reasons. First, particularly for bipolar transistors, the dc collector current that leads to minimum noise is considerably lower than that needed for maximum gain. Second, as we mentioned before, Γ_{MS} and Γ_{OPT} are usually far from each other, and we have to once again give up gain to achieve the best noise performance. This second reason also causes a poor input impedance match, leading to mismatch uncertainties when the amplifier is cascaded with other system components.

Two alternative approaches are available to overcome the fundamental problem of not achieving good input match when the active device is terminated with its optimum noise source impedance.

1. As long as we have sufficient gain, applying lossless feedback to the active device *may* bring Γ_{OPT} and Γ_{MS} closer together [14–16]. If the input mismatch loss is reduced to 0.5 dB or less,[9] the input match may be acceptable even when the source termination is selected for minimum noise. Lossless feedback also affects RF stability—it generally helps at the low gigahertz range but may cause problems at the higher microwave frequencies. A careful broadband stability analysis is highly recommended here.
2. Design the amplifier for minimum noise and use the balanced configuration [17] that also offers some redundancy if a device fails. This is a very practical approach for applications where the amplifier may be exposed to high voltage spikes, such as lightning, at an outside antenna. A balanced amplifier relies on the directivity of directional couplers to hide the poor input match of the LNAs. (The same concept may also be applied to high-power amplifiers where the low output impedance levels of the power transistors present problems.)

Let us look at examples for both techniques. For the first case, we use the BFP 640 already stabilized in Section 1.7.2.3, and also used for the

9. A 0.5-dB mismatch loss is equivalent to a 2:1 input VSWR that may be an acceptable specification for an LNA.

simultaneous conjugate match example in Section 2.2.1. Note that the device was characterized at a dc bias of 2V, 20 mA, and the *optimum condition for minimum noise* is at 2V, 5 mA. However, to preserve space we use the device already stabilized and leave the low-current design as an exercise for the reader. We need to emphasize, however, that the noise performance can be improved by several tenths of a decibel if the device is biased at 5-mA collector current instead of the 20 mA used in our maximum gain design example.

2.6.3 Illustrative example: design of a single-ended 1.9-GHz LNA

Figure 2.25 compares the optimum noise source reflection coefficient and the simultaneous conjugate-matched source terminations, superimposed with the available gain circles and the constant noise figure circles. Since Γ_{OPT} and Γ_{MS} are not very far from one another, at minimum noise condition we introduce relatively little mismatch loss at the input. Another way of stating this is that the input mismatch stays relatively low even when the active device is driven by a Γ_{OPT} source.

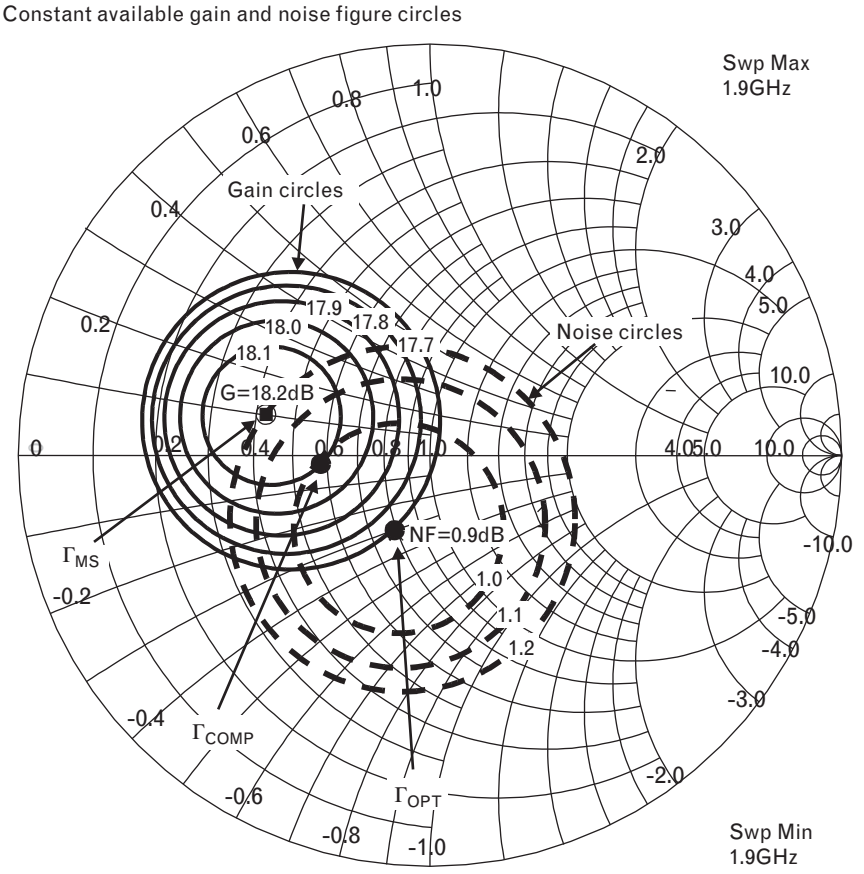


FIGURE 2.25
Γ_{OPT} and Γ_{MS} of the stabilized BFP 640 are relatively close to each other, indicating that we can expect reasonably good input match at minimum noise figure applications. Plotting available gain circles in 0.1-dB steps below G_{MAX} of 18.2 dB, and noise figure circles in 0.1-dB steps above NF_{MIN} of 0.9 dB, helps us to quickly arrive at a compromise solution at Γ_{COMP}. (Performance comparisons of the three options are given in Table 2.6.)

Viewing Figure 2.25, we have several options to design our LNA:

1. If the source is selected as Γ_{MS} for maximum gain, the corresponding noise figure is increased to 1.22 dB. Conjugate matching the output port to the load takes us back to the maximum gain amplifier of Section 2.2.1. In this case we see a perfect input match since there is no mismatch loss.
2. Operating the device with a source of Γ_{OPT} gives us minimum noise figure of 0.9 dB with 17.7-dB gain, which is 0.52 dB less than the maximum gain of the device. A 0.52-dB mismatch loss at the input converts to an input reflection coefficient magnitude of 0.34. This is not bad for an LNA, but we could do better by accepting a little higher noise figure.
3. Alternatively, we may choose a compromise source Γ_{COMP} between those two extremes and operate with higher than the minimum noise figure to improve the input match of the device. For instance, we can choose one-tenth of a decibel higher noise figure than NF_{MIN}, for which there is only 0.11-dB input mismatch loss, so that the input match is still very good.

Table 2.6 compares these three options. If we select the compromise source termination, Γ_{COMP}, we maintain good input impedance match with only a small sacrifice of noise figure. The available gain of the amplifier is expected to be 18.1 dB with Γ_{COMP} at the input. The output port should be conjugate-matched.

From (2.26), using the *S*-parameters of the stabilized BFP 640, we can compute the output reflection coefficient when the input of the device is terminated with $\Gamma_S = \Gamma_{COMP}$:

Table 2.6 Three Optional Source Selections for the Low-Noise Amplifier for the Stabilized BFP 640: Maximum Gain, Minimum Noise, and Compromise between the First Two

Source selected at	Γ_{MS}	Γ_{OPT}	Γ_{COMP}
Noise figure (dB)	1.22	0.9	1.0
Available gain (dB)	18.22	17.7	18.12
Mismatch loss at input (dB)	0.0	0.52	0.10
$\lvert s_{11} \rvert$ of the LNA	0.0	0.34	0.15
Input VSWR of the LNA	1.00	2.01	1.36

Note: The device parameters refer to a 2V, 20-mA dc bias condition that is *not* the optimum for low-noise operation. We used the same device parameters as before in the maximum gain amplifier.

$$\Gamma_{OUT} = s_{22} + \frac{s_{21}s_{12}\Gamma_{COMP}}{1 - s_{11}\Gamma_{COMP}} = (0.51\angle -1.7)$$
$$+ \frac{(6.64\angle 72°)(0.066\angle 67°)(0.25\angle -175°)}{1-(0.25\angle -92°)(0.25\angle -175°)}$$
$$= 0.61\angle -6.76°$$

Next, we compute the load to conjugate-match Γ_{OUT}:

$$\Gamma_L = \Gamma_{OUT}^{\star} = 0.61\angle 6.76°$$

For convenience, let us again use network topologies identical to the maximum gain amplifier of Figure 2.6. We can quickly determine the input and output circuit element values from the Smith chart. Figure 2.26 shows the schematic of the amplifier with the input circuit tuned for 1-dB noise figure. Once again, we are showing component values determined by graphical design, and all components are assumed to be ideal. Real-life modeling, layout, and statistical optimizations are not shown here,[10] but they should always be routine steps of practical circuit engineering.

Simulated performance (Figure 2.27) of the LNA verifies the accuracy of the available gain technique. At 1.9 GHz the amplifier has exactly 18.1-dB gain and 1.0-dB noise figure. The input reflection coefficient is less than 0.15, which is excellent for a low-noise amplifier.

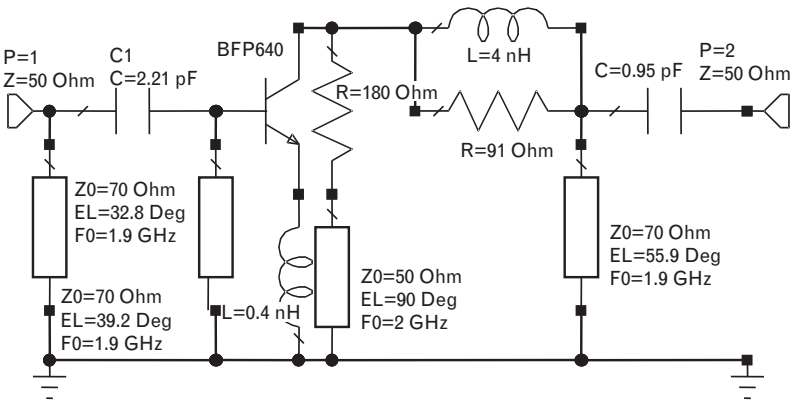


FIGURE 2.26 *RF circuit of the 1.9-GHz LNA, using the BFP 640 at 2V, 20-mA dc bias. The input circuit is tuned for a compromise solution between minimum noise figure and best input match. The output port is conjugate-matched to 50Ω. All element values are given for ideal components and require adjustments for physical realization.*

10. The physical layout of the parallel stabilizing branch was simulated in Section 1.10.

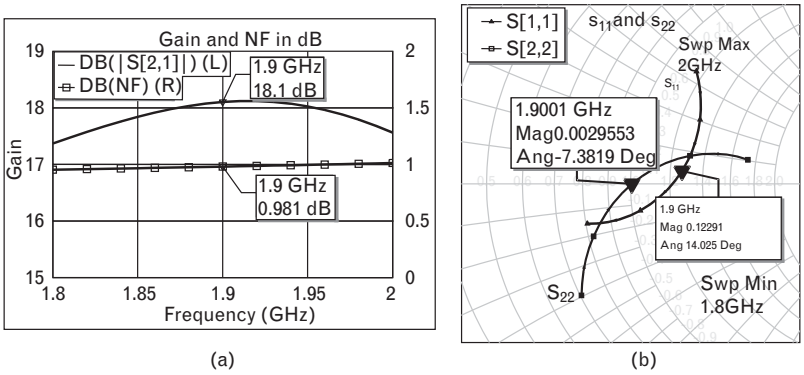


FIGURE 2.27 *(a) Gain and noise figure in decibels, and (b) input/output reflection coefficients of the 1.9-GHz low-noise amplifier. Improved noise performance can be achieved by biasing the device at a collector current of 5 mA, instead of the 20 mA used in our example.*

The output port of a narrowband low-noise amplifier can always be well matched. In this example we could also obtain simultaneously low noise and good input match due to the favorable characteristics of the active device. Other applications, particularly as the bandwidth requirements increase, may not achieve such good input match and we need to use an isolator or directional coupler to prevent excessive input mismatch. Our next example illustrates such a case.

2.6.4 Balanced amplifiers

Before we proceed with our second low-noise amplifier design, let us review the principles of *balanced amplifiers*. Figure 2.28 shows the topology of the balanced amplifier using two branch-line quadrature hybrid couplers and two identical single-ended amplifiers. One of the directional couplers splits the incident signal equally between the two amplifiers with 90° phase difference. The outputs of the amplifiers are summed by an identical coupler that also offsets the initial phase difference. The two active channels should be as symmetrical as possible, so circuit layout here requires special attention. For broadband applications the branch-line couplers may be replaced with more suitable structures, such as the *Lange coupler* [18].

One of the benefits provided by the directional couplers is that the individual amplifiers may have a poor impedance match that is not seen at the input and output ports of the balanced circuit. As long as the two channels are identical, signals reflected at their inputs or outputs are summed and dissipated in the 50-Ω resistors terminating the isolated ports of the couplers. As a result, the input and output ports of the balanced amplifier are matched to 50Ω and are completely isolated from the reflected signals. We also experience improved RF stability since the amplifiers are more isolated from outside terminations.

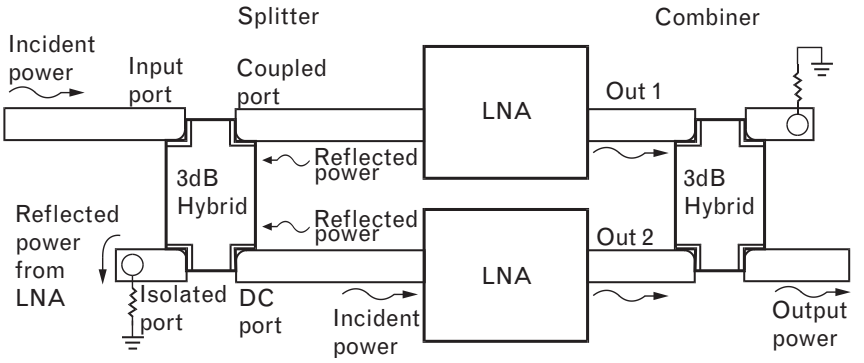


FIGURE 2.28 *Block diagram of a balanced amplifier using 90° hybrid directional couplers as power splitters and combiners. The two outputs of the splitter have equal magnitudes and 90° phase difference. Signals reflected from the two amplifiers cancel each other at the input port of the balanced amplifier but add in phase at the resistor terminating the fourth port. The output directional coupler performs similar tasks at the output side. (Courtesy of Agilent Technologies.)*

The balanced configuration has other advantages as well. First, for output power considerations, in power amplifiers two lower-power devices can be used to achieve the same output power as a single device. The lower-power devices are easier to match because their impedance is higher, and power dissipation is spread equally between them. Second, because of the symmetry of the balanced configuration, some odd-order spurious products are cancelled at the output (if they fall within the passband of the output coupler). The midpoint is not a virtual ground, because the two halves are not driven out of phase; thus, even harmonics do not cancel at the output. Third, the amplifier has a "soft-fail" mode, which is sometimes indispensable in systems requiring graceful degradation instead of sudden death. In the event of failure in one device, an alternative path still exists through the balanced amplifier. Although the VSWR and power will be degraded because the two arms are no longer balanced, the degradation allows for the fault and its location to be detected.

When we design low-noise amplifiers, since a directional coupler equally splits the incoming signal and noise, the noise figure of a single amplifier is also maintained in the balanced configuration, although the losses of the coupler increases the overall noise figure. The same applies to the output of a balanced power amplifier where coupler losses must be subtracted from the combined output power of the two channels.

Disadvantages of the balanced configuration include the expense of the additional second amplifier chain, the finite insertion losses of the two couplers, and the need to maintain the phase and amplitude matching of the two channels. In some circumstances, however, using a balanced configuration might be the only solution. For example, at millimeter wave frequencies large output powers can only be achieved by summing the power from multiple devices.

To summarize, balanced amplifiers nearly have the gain of each individual amplifier. Noise figures of the individual amplifiers are maintained while their output powers are summed. However, coupler losses reduce the gain and output power and also increase noise figure. For example, assume that each of the two amplifiers in Figure 2.28 has 40-dB gain, 2-dB noise figure, and maximum linear output power of +30 dBm (1W). If we use 0.2 dB for the individual coupler losses, the balanced amplifier has 2.2-dB noise figure, 39.6-dB gain, and maximum linear output power of +32.8 dBm, or 1.9W.

2.6.5 Illustrative example: design of a balanced LNA for the 1.7- to 2.3-GHz frequency range

We use the Agilent *enhancement mode* HEMT (AFT-54143) to illustrate a balanced low-noise amplifier design. The FET is biased at $V_{DS} = 3.0$V at $I_D = 60$ mA. One of the advantages the FET has over bipolar devices is the outstanding noise performance even at relatively high drain currents. Low noise combined with high output power leads to increased dynamic range. Additional details, including measured nonlinear performance are available in [19, 20].

Enhanced mode FETs are closer to bipolar transistors than *depletion mode* types in the sense that the device is turned off when the gate is zero biased.[11] Accordingly, we can bias the AFT-54143 from a single dc source, with either a passive or one of several active circuit arrangements, such as a current-mirror [21] circuit.

As for most of the RF transistors, common-ground feedback significantly affects the RF parameters of the AFT-54143. A small amount of inductive reactance can help RF stability at some frequencies and may hurt at others. Before we proceed with the circuit design, we need to evaluate the RF grounding path and include it in our computations. In this case, the device parameters were measured on a 25-mil-thick (0.635 mm) PC board using via-hole grounding, and here we need to modify the parameters to simulate the performance on a thicker board. The device's package has two source contacts and we include both of them in our simulations.

Two short traces of conductors grounded through four via holes amount to 0.45-nH equivalent inductance. We use that inductance in this example to compare source terminations for maximum gain and minimum noise in the 1.7- to 2.3-GHz band.

If the emitter of the AFT-54143 is directly grounded, the device is potentially unstable at 1.8 GHz; therefore, the magnitude of Γ_{MS} is greater than unity. Adding 0.25- to 0.3-nH inductance into the common (emitter) terminal stabilizes the device and brings Γ_{MS} inside the Smith chart. The

11. Actually, even with zero bias on the gate there is a small drain current due to drain-gate leakage.

conductor traces used in the layout of this amplifier amount to about 0.45-nH inductance, providing greater than unity stability factor. In addition to helping stability, the ground inductance also brings Γ_{OPT} and Γ_{MS} closer to each other.

Figure 2.29 shows that around 1.8 GHz, even with the added source inductance, Γ_{OPT} and Γ_{MS} are still separated by nearly 1 dB of gain. An amplifier tuned for minimum noise would have poor input match in a 50-Ω system. Additional lossless feedback to bring Γ_{OPT} and Γ_{MS} closer together leads to lower gain and potential stability problems at higher frequencies. However, using balanced configuration with two parallel LNAs, the directional couplers "hide" the mismatch, and we can have low noise and good impedance match simultaneously.

Once again, using the available gain approach we design a low-noise amplifier in a 50-Ω system. A two-element highpass network transforms our 50-Ω source termination to Γ_{OPT} of the device, and another two-element highpass section matches the resultant output impedance to the 50-Ω load. Adding a passive dc bias network with RF filtering completes one of the two amplifier channels to be used in the balanced amplifier.

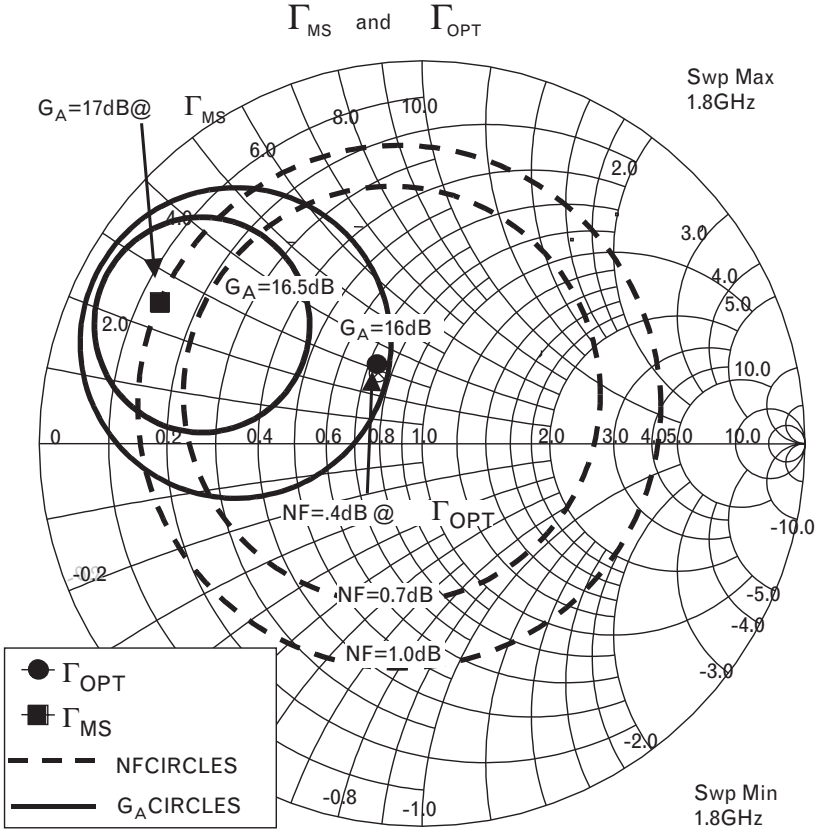


FIGURE 2.29
At 1.8 GHz, Γ_{OPT} and Γ_{MS} of the inductively grounded AFT-54143 are separated by almost 1-dB gain and 0.6-dB noise figure. Choosing the source at Γ_{MS} (marker), we get 17-dB gain, perfect input match, and 1-dB noise figure. Having the source at Γ_{OPT} (marker) leads to minimum noise figure of 0.4 dB, slightly over 16-dB gain, and about 1-dB mismatch loss. With a few tenths of a decibel increase in noise figure, the balanced amplifier arrangement brings low noise and good match together.

Initial circuit design with ideal components (Figure 2.30) provides the starting point for the more detailed second phase that takes us to the real-life circuit. Actual final component values depend on component parasitics and circuit layout.

The dc biasing of the FET is controlled with the passive network of resistors R_2 through R_5. Drain voltage is reduced to 3V from the 5V dc supply by resistor R_5. Gate voltage is set to 0.56V with the voltage divider of R_3 and R_4. Resistor R_2 provides current limiting to the gate in case the device is overdriven by a large input signal. Resistor R_1 presents a resistive source to the device for RF stability at low frequencies.

At the input side of the FET, the three-element LC network L_1-C_1-L_2 transforms the 50-Ω output impedance of the directional coupler to Γ_{OPT} of the device. The dc bias voltage is applied to the gate of the FET through inductor L_2. Capacitor C_1 also serves for dc blocking. At the output side, L_3-C_4 provides impedance matching as well as biasing and blocking. Capacitors C_2 and C_5 present RF shorts to the matching circuits, while elements LL_1 and LL_2 form the total grounding inductance of the FET. The purpose of C_3 and C_6 is low-frequency decoupling.

The two LNAs are connected together with a miniature surface-mount hybrid coupler that was specially designed for the 2-GHz frequency band. Some 3-dB hybrids maintain amplitude and phase balance over a 20% to 30% bandwidth [22]. For example, the Anaren JP503 Pico Xinger coupler's outputs have typical amplitude balance of ±0.2 dB and phase balance of ±2° at 0.25-dB insertion loss between 1.9 and 2.5 GHz. The coupler is characterized by the four-port *S*-parameters provided by the manufacturer.

Symmetry has vital importance in the layout of a balanced amplifier. Unless we maintain identical RF path lengths between the two channels, the desired signal summations and reflection cancellations do not take place. Since active and passive component tolerances also play important roles in the overall performance, the individual amplifiers of the balanced

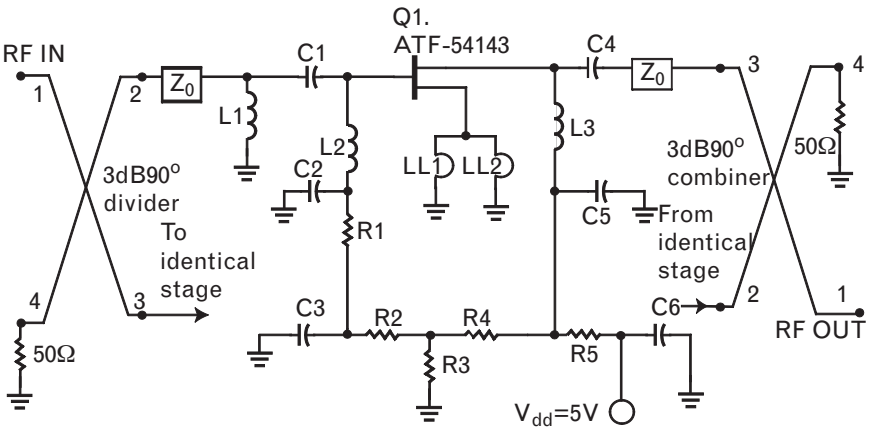


FIGURE 2.30
Circuit schematic of the LNA shows initial RF component values and a passive dc bias network. (After [20].)

circuits are ideal candidates for RFIC realizations. In discrete circuit form, we need to hold tight component tolerances and, if possible, build circuits with components coming from single production lots.

Once the initial circuit layout is formed, we need to resimulate the circuit, including the actual component models, circuit parasitics, and discontinuities. Most likely this step requires at least one more circuit optimization and may even require the assistance of EM simulation. Remember that the time invested in this phase can pay off a hundredfold later by preventing production problems.

Figure 2.31 shows the final detailed circuit schematics of a single LNA channel. Each of the passive surface-mount type passive components is modeled by their physical equivalent circuit models that include losses and parasitics. Microstrip transmission line discontinuities, such as a 90° bend or a T-junction, are also included by their corresponding two- or three-port models. RF and dc grounds are achieved with via-hole models. The FET is represented by its two-port *S*-parameters.

The final physical circuit (Figure 2.32) was realized on 31-mil-thick FR-4 dielectric board using surface-mount component technology. The

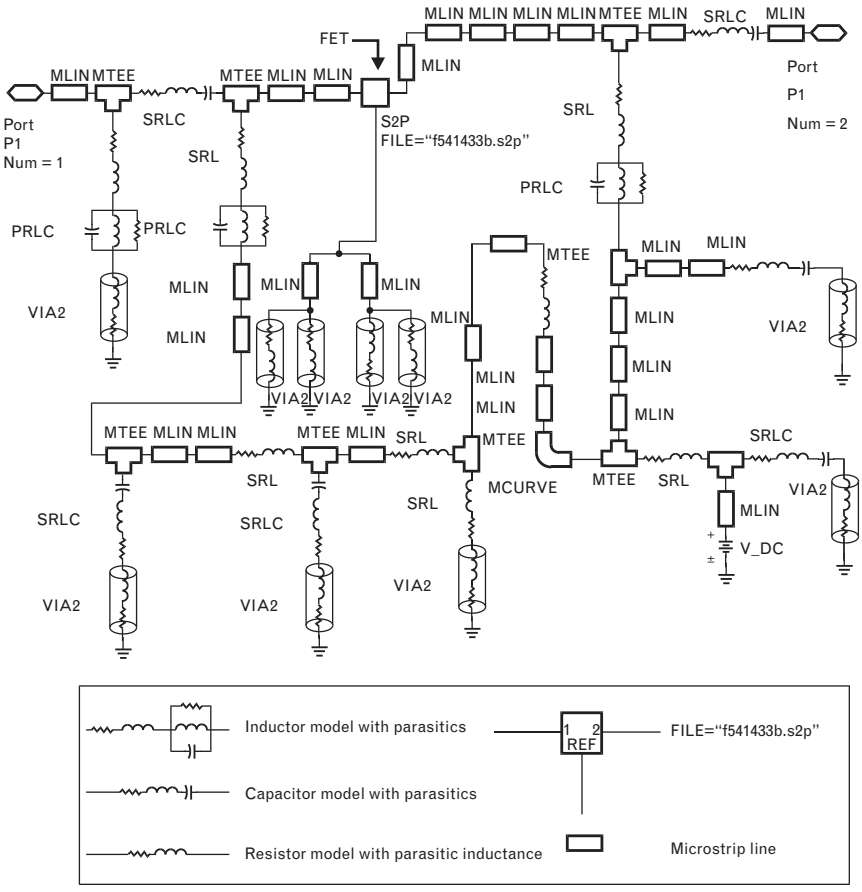


FIGURE 2.31
ADS schematics of one of the two identical LNAs. The two-port S-parameters of the active device may be replaced with the appropriate equivalent circuit model for nonlinear simulation. (Source: [20]. © 2002 Agilent Technologies. Reprinted with permission.)

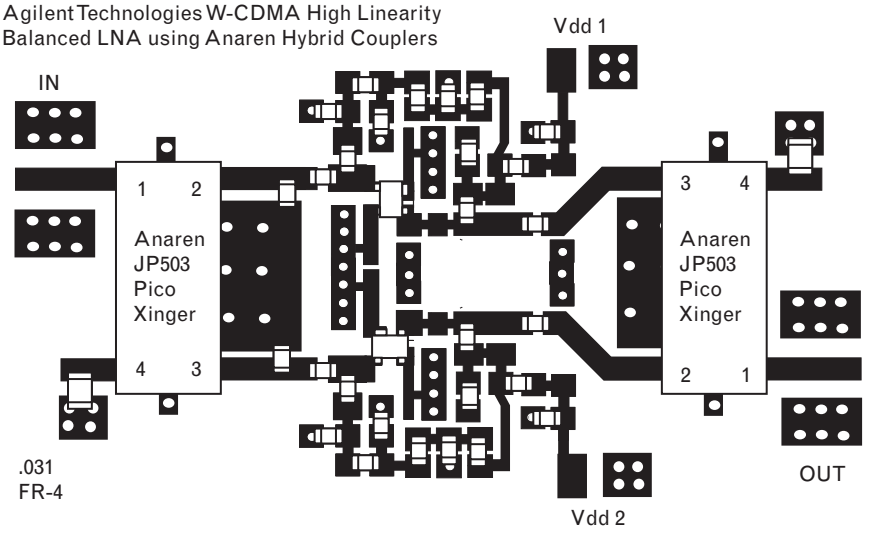


FIGURE 2.32
Layout of the balanced amplifier requires special care to keep the two LNA channels' signal paths identical. The 3-dB hybrids couplers split and combine the RF signal with 90° phase difference. (After [20].)

50-Ω microstrip transmission lines located between grounded strips provide RF input and output connections. A single dc power supply is connected between ground and the terminals marked V_{dd1} and V_{dd2}. The PC board was laid out to provide variable ground inductance to the two source terminals of the FET, since the same board is also used for other amplifiers operating at different frequencies.

Due to the high degree of modeling, simulated and measured performances of the balanced amplifier are very close. Figure 2.33 shows the performance through the 0.5- to 3.0-GHz frequency range. Input and

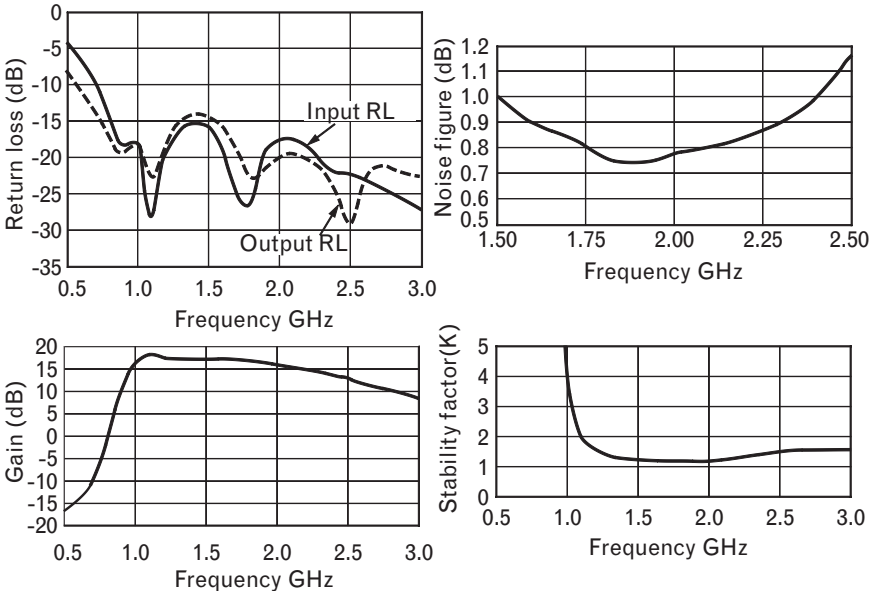


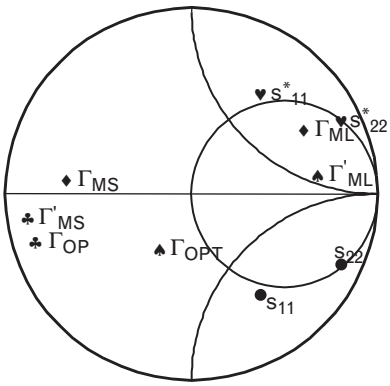
FIGURE 2.33
Simulated input/output return loss, gain, noise figure, and RF stability of the complete balanced amplifier, including both directional couplers. (Source [20]. © 2002 Agilent Technologies. Reprinted with permission.)

output return losses are better than 17 dB, gain is 16 ± 1.0 dB, and noise figure is less than 0.85 dB. The 1-dB gain compression point is +22.5 dBm (nearly 200 mW) at 2.0 GHz, assuring a very wide dynamic range. Total dc current consumption of the amplifier is 120 mA at +5.0-V power supply.

2.7 Comparison of the various amplifier designs and Smith chart–based graphical design aids

S-parameter–based design techniques enable us to design linear RF amplifiers for several different applications. Maximum small-signal gain comes with simultaneously conjugate-matched ports that they are readily cascadeable with other matched system blocks. Amplifiers designed for maximum absolute linear output power have a mismatched output port and a matched input. A low-noise amplifier is just the opposite: it has a matched output and mismatched input ports. The last two categories may present problems when cascaded with other system components, particularly in broadband applications where mismatch uncertainties can cause serious gain ripple.

Figure 2.34 compares the terminations required for the three amplifier types, as well as the terminations used for unilateral maximum gain. It is clear from the plot what important roles the terminations have in *S*-parameter design. Γ_{MS} and Γ_{ML} are the simultaneously matched terminations. Γ_{OPT} and Γ_{OL} are the optimum terminations for minimum noise and maximum output power, respectively. However, in these two cases only



Source & load combinations for:

Symbol	Amplifier Type	Source	Load
♥	G_{UMAX}	s^*_{11}	s^*_{22}
♦	G_{MAX}	Γ_{MS}	Γ_{ML}
♠	LNA	Γ_{OPT}	Γ'_{ML}
♣	Power	Γ'_{MS}	Γ_{OL}

FIGURE 2.34 *Comparison of source and load terminations of four S-parameter design techniques for three types of linear RF amplifiers. Parameters s_{11} and s_{22} are the reflection coefficients of the active two-port. For maximum gain amplifiers we show two ways: one with unilateral (G_{UMAX}) and another with bilateral (G_{MAX}) approach. Γ'_{ML} and Γ'_{MS} are the matched source and load terminations used in available and operating gain design.*

the adjacent port is matched, labeled as Γ'_{ML} and Γ'_{MS}, respectively, to differentiate from the simultaneous conjugate match.[12] Graphical design aids, such as the Smith chart, help us to determine the appropriate source and load terminations and transform the existing system terminations to the specified ones.

Let us now summarize the three major amplifier categories based on the order of the terminations determined.

1. *Maximum small signal gain:*

   Unilateral design: (Not recommended, unless the *U*-factor is very low.) Use s^*_{11} and s^*_{22} for source and load terminations.

   Bilateral design: Find Γ_{MS} and Γ_{ML}, simultaneously. For unconditionally stable two-ports, Γ_{MS} and Γ_{ML} are always inside the Smith chart. No practical solutions exist for potentially unstable two-ports.

2. *Maximum linear output power:* Use the operating gain expression. Terminate the output port with Γ_{OL} and compute the new input reflection coefficient, Γ_{IN}. Provide a source termination $\Gamma'_{MS} = \Gamma^*_{IN}$. This technique works regardless of the stability of the two-port, although the resulting output reflection coefficient magnitude may be very high for potentially unstable two-ports.

3. *Minimum noise:* Use the available gain expression. Terminate the input port with Γ_{OPT} and compute the new output reflection coefficient, Γ_{OUT}. Provide a load termination $\Gamma'_{MS} = \Gamma^*_{OUT}$. This technique works regardless of the stability of the two-port, although the resulting input reflection coefficient magnitude may be very high for potentially unstable two-ports.

Figure 2.35 compares the three bilateral design techniques in their generalized forms. Remember that we need to obtain two terminations for any amplifier to be designed. With a set of given two-port *S*-parameters, the transducer gain is a function of two terminations—and we cannot solve a single equation with two unknowns. We need to write two equations to obtain the unique set of simultaneously matched terminations, Γ_{MS} and Γ_{ML}.

Operating and available gain expressions have one unknown termination that is first selected for a special task, such as maximum linear output power or minimum noise. Only then can we compute the second termination to match the adjacent port. Under those conditions the gain of the amplifier is less than the maximum gain of the two-port, because only one of the two ports is impedance matched.

12. In the available and operating gain procedures, only one port gets matched termination, marked with Γ'_{ML} and Γ'_{MS}, respectively.

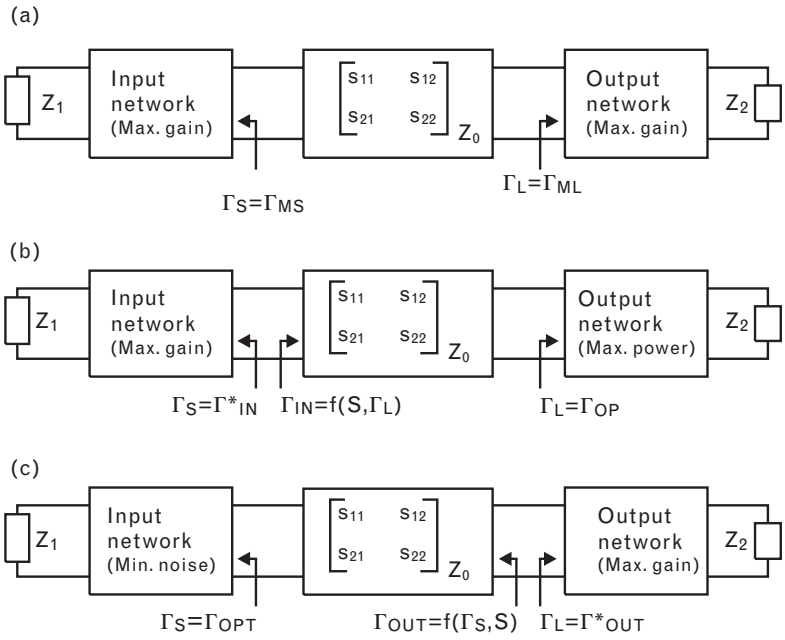


FIGURE 2.35
Generalized block diagram of the three sets of amplifier matching networks for (a) maximum gain, (b) maximum linear output power, and (c) minimum noise. Performance is a function of the two-port's parameters and terminations.

Table 2.7 summarizes the definitions and functions of the various Smith chart design aids defined both in this volume and in Volume I. Although most of the RF/MW circuit simulation programs readily plot these circles and contours, for reference we include the related mathematical formulas in the appendix.

2.8 Broadband amplifiers

Although there are no set rules to consider, an amplifier is generally considered to be narrowband when its bandwidth is less than 20% of the center frequency. Broadband amplifiers, on the other hand, can cover extremely wide bandwidths. Amplifiers used in military defense systems and test equipments often require multidecade frequency range coverage. For example, a network analyzer covering from kilohertz through gigahertz range operate through more than five decades of frequencies. It is impossible to rely on reactive elements for impedance matching for such a wide frequency range. If a series capacitor's capacitive reactance is $-j30\Omega$ at 1 GHz, it increases to $-j3\text{M}\Omega$ at 10 kHz, which is not very practical for impedance matching.

As we saw earlier, single-section matching networks in amplifiers can generally cover 10% to 15% fractional bandwidth easily, except when (1) the real parts of the termination ratios approach or exceed about 5, and/or (2) the Q's of the required terminations are greater than 2 or 3.

Table 2.7 Useful Amplifier Design Aids Based on Small-Signal Two-Port *S*-Parameters, Assuming Source and Load Terminations Selected within the Smith Chart

Name	Descriptions	Value limits	Section
Constant *Q*-circles	Arcs connecting all points of equal *Q*'s. Mirror-image symmetry between upper and lower half of the Smith chart	0 to ∞	Volume I, 4.6
Unilateral constant-gain circles of sources	Circumference is the locus of all source terminations that change the gain by the same amount. Does not include any interaction with the load	$-\infty$ dB to G_{1MAX} dB	1.4.3
Unilateral constant-gain circles of loads	Circumference is the locus of all load terminations that change the gain by the same amount. Does not include any interaction with the source	$-\infty$ dB to G_{2MAX} dB	1.4.3
Source stability circles	Locus of all source terminations that lead to $\lvert\Gamma_{OUT}\rvert = 1.0$. Includes the effect of input-output interaction	N/A	1.5.4
Load stability circles	Locus of all load terminations that lead to $\lvert\Gamma_{IN}\rvert = 1.0$. Includes the effect of input-output interaction	N/A	1.5.4
Operating gain circles	Locus of all load terminations that lead to the same overall operating gain. The input port should be conjugate matched. Includes the effect of input-output interaction	$-\infty$ dB to G_{MAX} dB for $\mu > 1$ $-\infty$ dB to MSG dB for $\mu < 1$	2.4
Available gain circles	Locus of all source terminations that lead to the same overall available gain. The output port should be conjugate-matched. Includes the effect of input-output interaction	$-\infty$ dB to G_{MAX} dB for $\mu > 1$ $-\infty$ dB to MSG dB for $\mu < 1$	2.6
Constant output-power contours	Locus of all load termination that lead to the same absolute output power. May be *approximated* by sections of constant-resistance and constant conductance circles	$-\infty$ dB to P_{MAX} dB	2.4.4 and 5.2.2

Increasing the order of the matching networks generally helps, unless the input or output termination *Q*'s pose fundamental gain-bandwidth limitations. In that case we can only get improvement by reducing the primary parasitics. Using better packages for the active devices or switching to chip technology may be the way to obtain wider bandwidth.

In this section, we discuss the design of amplifiers for a wide frequency range. The four most commonly used broadband techniques rely on selective gain equalization and negative feedback.

2.8.1 Reactive match/mismatch approach

With the help of constant-gain circles, we can selectively increase or decrease the basic transducer gain of the active device, as shown in Figure 2.36 [23]. If our goal is to cover a frequency band between f_1 and f_2,

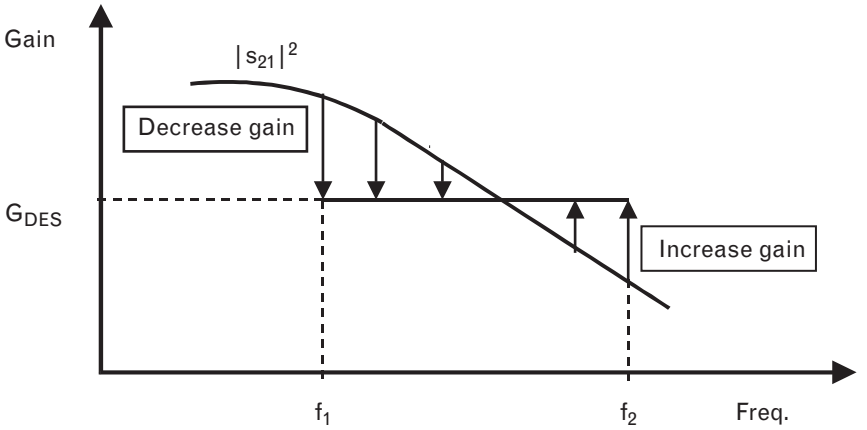


FIGURE 2.36
The basic 50-Ω gain of a two-port may be flattened between frequencies f_1 and f_2, by selectively applying match or mismatch. Constant-gain circles can be helpful to find the necessary reactive circuit elements.

the gain at the highest frequency should be 1 to 2 dB below G_{MAX}. Trying to get the maximum gain at f_2, leads only to a single set of source and load terminations at that frequency: Γ_{MS} and Γ_{ML}. It is unlikely, however, that Γ_{MS} and Γ_{ML} also satisfy the broadband gain requirements. Setting a lower goal for the desired gain, G_{DES}, the source and load terminations lie on specified gain circles, offering more circuit options and better chance for a broadband solution.

The fundamental weakness of this approach is poor impedance match—by throwing away gain we create reflection. Using lossless matching circuits, *what is not transmitted must be reflected*. It is easy to see that if we compensate a 6-dB gain roll-off at one port, the maximum resultant reflection coefficient magnitude is 0.87—which is totally unacceptable for cascading the amplifier to other system blocks. If the gain compensation is applied on both sides of a two-port, we also expect uncertainties since we are using an approach based on the unilateral assumption. Circuit optimization can be helpful to finalize component values.

In multistage amplifiers the gain compensation may be selectively distributed throughout the interstage networks to avoid poor input and output match. Still, in most broadband applications, the amplifier must be placed between directional couplers or isolators to hide the poor impedance match from other components.

Practical component realizations limit the maximum bandwidth of this approach to one to three octaves of frequency in most cases, depending on the Q's of the device impedances. Impedance transformation ratios between the device and the terminations are also limiting. The same constraints also apply to the two lossy matching techniques described next.

2.8.2 Dissipative mismatch at input and/or output ports

Since we must give up some gain at the lower frequency, the unwanted gain could be dissipated instead of being reflected. This approach provides

a better impedance match at the input and output of the amplifier. Figure 2.37 shows two forms of such dissipative branches where the added resistors dissipate more signal power at lower frequencies. By making a careful choice between series or parallel resistance, we can increase or decrease the impedance of the applicable port and improve impedance matching.

Let us restate that we are generally against using resistors for impedance matching since they do not really match—they just cover up mismatch. In broadband amplifiers, however, the active devices have more than the desired gain at lower frequencies. It is better to dissipate than to reflect the excessive gain, since we may also be able to lower the port reflection coefficient this way.

2.8.2.1 Illustrative exercise: single-stage 800- to 2,000-MHz broadband amplifier

To illustrate the dissipative match-mismatch approach, we again use the BFP 640 bipolar device that was already stabilized for all frequencies in Section 1.7. Our goal is to have an amplifier that works from 800 through 2,000 MHz, a 2.5:1 frequency ratio, with good input/output match and flat gain response.

Figure 2.38 compares the 50-Ω gain and maximum gain of the stabilized device, showing that G_{MAX} drops from 22.3 to 18 dB through the desired frequency range. The 50-Ω gain rolls down from 22.3 to 16.3 dB, having a value of 18 dB at 1,339 MHz.

If we perfectly match both ports of the device throughout the whole frequency range, we end up with 5.3-dB down-slope. On the other hand, if we progressively reflect the unwanted gain, the mismatch losses cause very poor impedance match. For example, at 800 MHz we need to mismatch the device by about 8 dB. Even if we split the mismatch between the two ports, a 4-dB mismatch loss converts to 0.78 reflection coefficient magnitude, which is unacceptable for system applications.

The only way to maintain flat gain and good impedance match is to introduce lossy frequency selective gain shaping. By sacrificing 3- to 4-dB

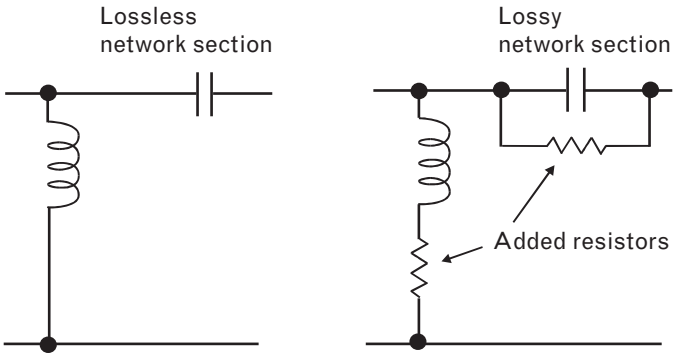


FIGURE 2.37
Losses may be added to LC matching section to dissipate unwanted gain instead of reflecting it. In both configurations the resistors absorb more signal power at lower frequencies.

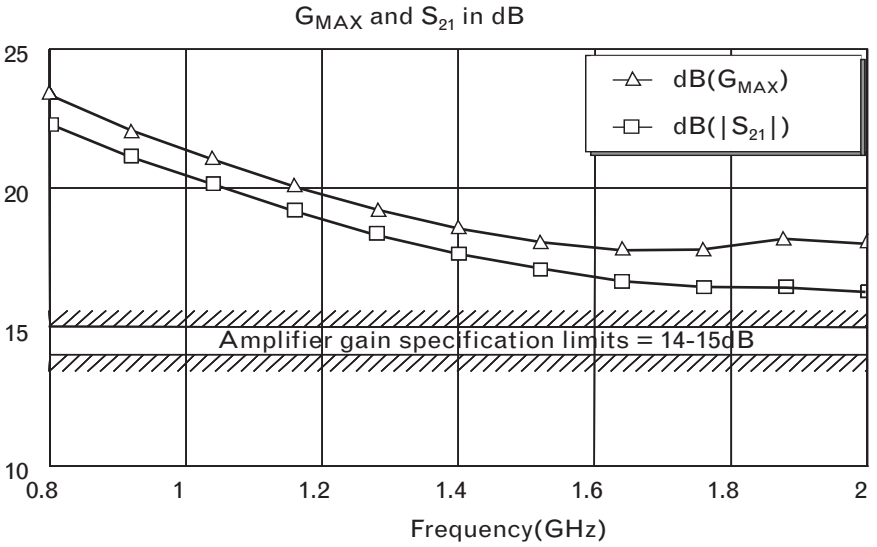
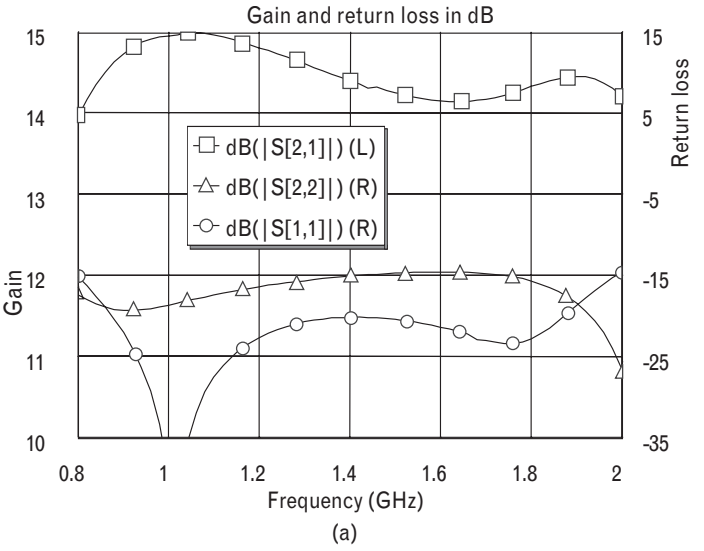


FIGURE 2.38 *G_{MAX} and $|s_{21}|_{dB}$ of the stabilized BFP 640 roll off nearly 6 dB/octave between 800 and 1,600 MHz. The gain roll-off slows at 1,600 MHz as the stability factor of the device begins to decrease. G_{MAX} actually shows a slight increase between 1,800 and 1,900 MHz, due to the quarter-wave effect of the short-circuited parallel stub used the stabilizing network. Shaded region shows the 14- to 15-dB target gain of the amplifier.*

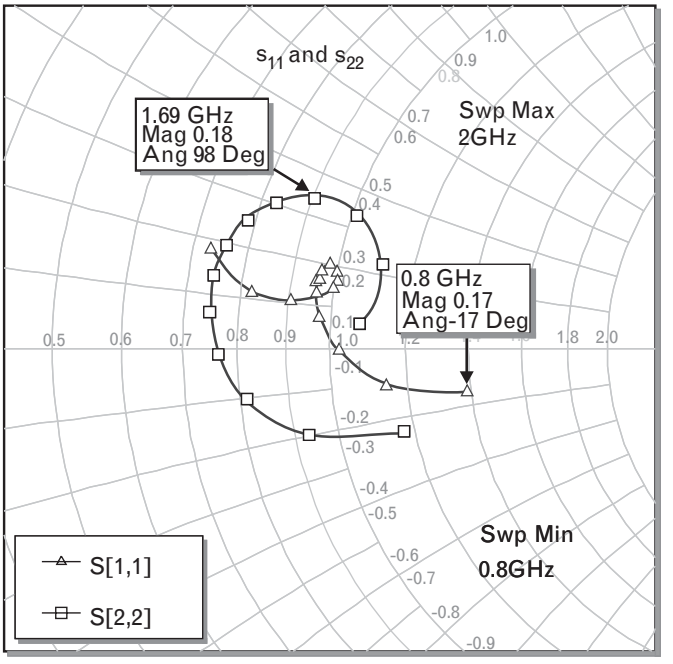
gain at the high end of the bandwidth, we can maintain flat frequency response and good input output match for the whole frequency range. In our example we target the broadband gain to be between 14 and 15 dB for the 800- to 2,000-MHz frequency range. The goal is to have a well-matched amplifier to cover both the low and high frequency bands of cellular telephone communications.

To cover the broad bandwidth with selective gain compensation and maximum symmetry (see Volume I, Chapter 5), we chose fourth order matching networks on both sides of the device. Initially, the matching networks were designed at the geometric band center frequency (1,265 MHz) by using single-section lowpass and highpass circuits at both sides. After the two fourth-order lossless matching networks were derived, we added a small amount of resistance in series with each inductor. Since the input impedance of the stabilized device is lower than its output, the input circuit needed series loss. On the output side, parallel loss improved the match to 50Ω. We then submitted the circuits to optimization, targeting flat gain and low input/output reflection coefficients.

Circuit optimization gave us good results in the upper 60% of the bandwidth but had problems with the impedance match at low frequencies. Both the input and output ports were still too capacitive. Adding one more parallel inductor to the input and output ports and reoptimizing the new circuit gave us very good results, as shown in Figure 2.39.



(a)



(b)

FIGURE 2.39 *Frequency response and impedance match of the 800- to 2,000-MHz amplifier. (a) Gain is 14.5 ± 0.5 dB on the left-side scale, and return loss is better than 14 dB on the right scale. (b) Input and output reflection coefficients are plotted on the expanded Smith chart, showing magnitudes less than 0.2. Markers indicate the highest magnitudes of s_{11} and s_{22}.*

The RF circuit schematic of the final optimized amplifier is shown in Figure 2.40. All components are assumed to be ideal. Gain flatness is within ±0.5 dB, and both input and output reflection coefficient magnitudes are less than 0.2 throughout the 800- to 2,000-MHz frequency range

How the cost and performance of a discrete broadband amplifier compare to RFICs is up to the system designer to decide. Prices of RFICs have

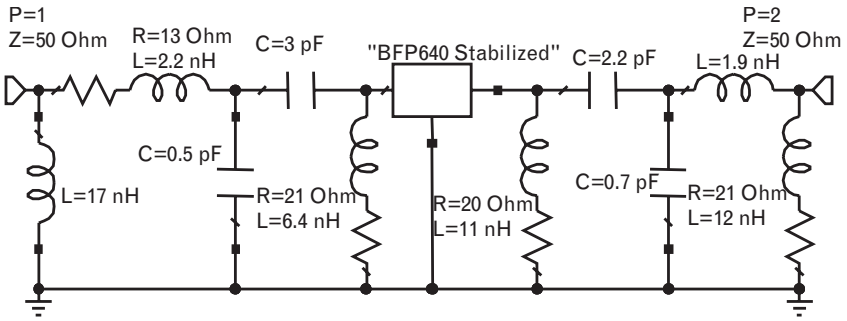


FIGURE 2.40 *RF circuit schematics of the broadband amplifier. If the series RL combinations at the input and output sides of the device are RF-grounded, dc bias may be applied to transistor, with capacitors C_2 and C_3 providing dc blocking.*

been coming down while their performances have been improving. Levels of integration have been simultaneously increasing, and RFIC building blocks already dominate most of the low-cost RF communication systems.

2.8.3 Amplifier-equalizer combinations

If we are able to match our active device to a reasonable level ($VSWR < 2.0$), then cascading matched gain equalizers is another way to obtain flat overall gain. Such an active gain block has an approximate 6 dB/octave gain roll-off. A *bridged-T* or similar passive gain equalizer [24] offers good input and output impedance match with a positive gain slope. Cascading the matched equalizer to a semi-matched amplifier and applying computer optimization can bring flat gain and good impedance match. Figure 2.41 illustrates such a cascaded arrangement.

2.8.4 Feedback amplifiers

Instead of reflecting or dissipating the unwanted low-frequency gain, we can apply negative feedback to the active device. When properly designed, negative feedback can:

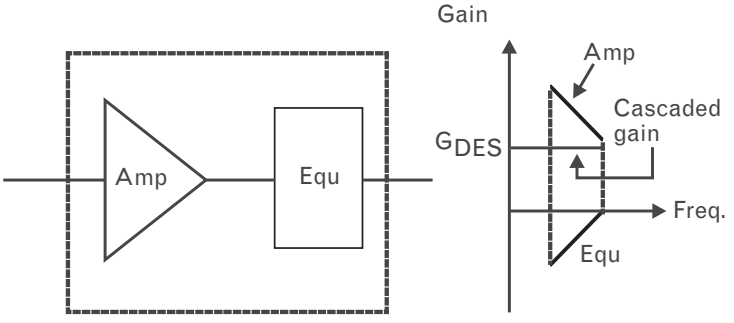


FIGURE 2.41 *Cascading an amplifier (AMP) and equalizer (Equ) results in flat gain if both two-ports are matched. Ripples caused by mismatches may be reduced by optimization.*

- Maintain gain flatness and impedance match;
- Reduce temperature component tolerance effects;
- Improve dc and RF stability;
- Reduce distortion (depending on where the distortion is created) [25].

Positive feedback does just the opposite, so we need to be extremely careful about the phase relationship between the input signal and the signal being fed back from the output. Applying overall feedback through cascaded stages is very difficult at RF and should never be done without a detailed computer-aided analysis.

The common-emitter configuration of a bipolar transistor and the common-source configuration of an FET provide 180° phase difference between the input and output at dc; therefore, they are ideal candidates for negative feedback applications. Common-base and common-collector arrangements of bipolar transistors, and their equivalences for FETs, are not useful for that purpose because their outputs are in-phase with the inputs.

As the frequency increases, the parasitic elements of the active devices cause an additional phase shift (Figure 2.42) and even the common-emitter configuration of a bipolar transistor can lead to positive feedback. When the phase of s_{21} becomes less than 90°, any signal fed back from the output to the input, purposely or unintentionally, has as an in-phase component

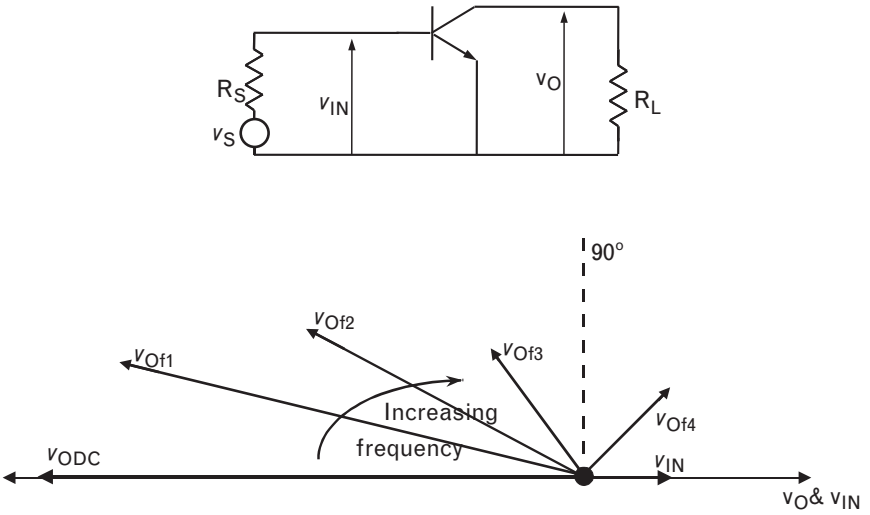


FIGURE 2.42 *Input-output signal voltage relationship of a bipolar transistor in common-emitter configuration at various frequencies. The magnitude of output voltage gradually decreases from dc (v_{Odc}) through frequency f_4 (v_{Of4}) and the phase is gradually shifting in a clockwise direction. After the output voltage phase becomes less than 90°, its real part is in phase with the input voltage.*

that increases the input level and we no longer have negative feedback. Of course, an open-loop examination of the *S*-parameters measured in the 50-Ω system does not predict accurately the closed-loop performance, but it can still provide a warning sign of possible problems.

Generally speaking, parallel feedback alone helps stability at lower RF frequencies but weakens it at the higher frequencies. Series feedback can decrease stability much sooner than parallel feedback, and it is not recommended for RF applications. However, if we want flat broadband gain and good impedance match simultaneously in the RF range, we need to apply both types of feedback.

A detailed loop transmission analysis of the feedback amplifier circuitry [26] is beyond the scope of this chapter because it requires an accurate equivalent circuit for the active device. Simplified models have been used to approximate the performance [2], but they do not include the effects of internal feedback and parasitics.

Although the open-loop gain-phase characteristics change with external feedback and different terminations, it is still useful to know the frequency range where the phase of s_{21} crosses 90°, which is the border between negative and positive feedback. The effect of the *internal feedback* of the device (such as the Miller capacitance of a bipolar transistor) is inversely proportional to the load. Since transistors are not conjugate-matched in feedback amplifiers, the effective loading of a device is generally less than 50Ω. Therefore, the measured s_{21} of a device in a 50-Ω system can give us a conservative estimate of the upper frequency limit for negative feedback.

Figure 2.43 shows the change of s_{21} of an Infineon BFP 520 transistor characterized in a 50-Ω system. For this device the 90° open-loop phase crossover is just above 1.8 GHz. Adding external feedback moves that frequency considerably higher, and we will later use this device in a feedback amplifier design where the goal is to have flat gain to 4 GHz.

External passive feedback circuits increase the magnitude of s_{12} and result in less isolation between the input and output. Remember that these amplifiers are definitely not unilateral. Any circuit adjustment on one side of a feedback amplifier has strong effect on the other side. Impedance matching should always be performed with the bilateral approach covered in Section 2.2.

As long as the difference between the basic open-loop gain and the desired amplifier gain is at least 5 to 10 dB, negative feedback helps to maintain flat gain. Combinations of series and parallel feedback may also control the input and output impedances. Reaching the frequency where the gain of the feedback amplifier is equal to the basic gain of the two-port (f_2 of Figure 2.44), the amplifier's gain starts to roll off. Remember that if we truly have negative feedback, the *amplifier's gain cannot exceed the basic 50-Ω gain of the active device*. However, at frequency f_2 the device is still

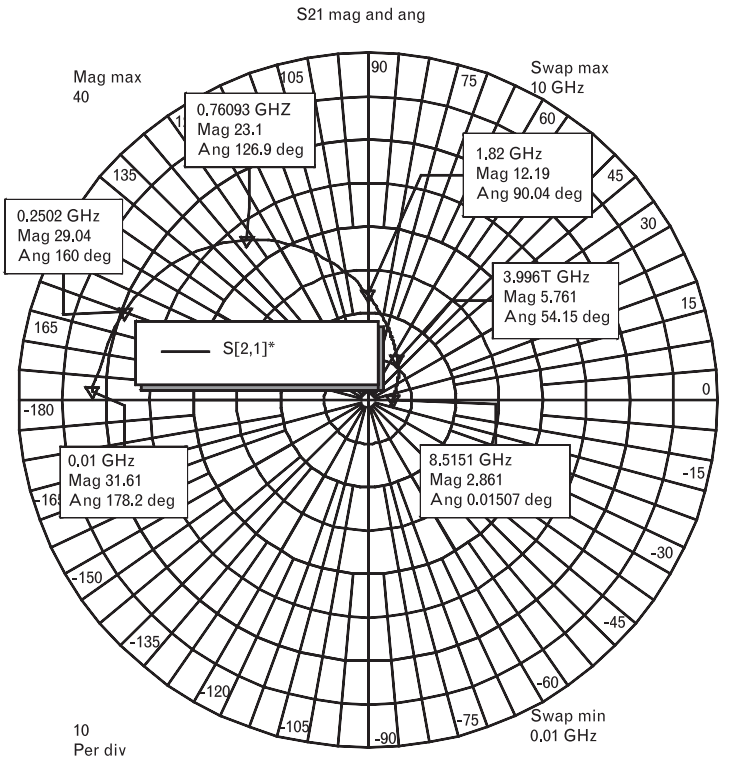


FIGURE 2.43
Plotting s_{21} of a BFP 520 bipolar transistor on a polar chart shows a gradual decrease of the magnitude and an additional phase-shift with the increase of frequency. At 0.01GHz, the magnitude is 31.61 (30 dB) with nearly perfect 180° phase shift. As the frequency increases to 1.821 GHz, the phase reaches 90° and the magnitude drops to 12.19 (21.7 dB). At 8.516 GHz the device still has about 9.1-dB gain and the output signal is exactly in-phase with the input.

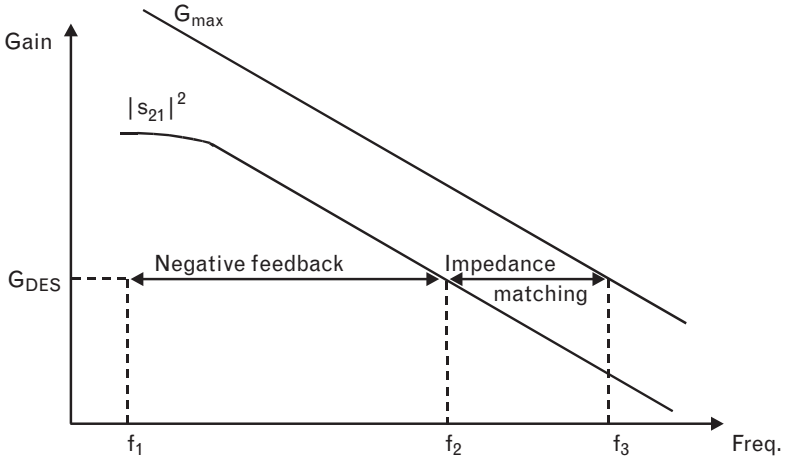


FIGURE 2.44
Negative feedback cuts the open-loop gain of the two-port to G_{DES} and maintains flat gain response from f_1 to nearly f_2. Above f_2, impedance matching with lowpass networks may be used to extend the amplifier's frequency range to f_3.

capable of more gain if we apply impedance matching. Theoretically, without component losses, with gradual phase-in of impedance matching the amplifier's gain level, G_{DES}, could be extended to frequency f_3.

We saw earlier in the narrowband design that using highpass matching network topologies offers several benefits, one of which was to cut the unwanted low-frequency gain. Well, those networks certainly do not work here, because the last thing we want is to change the already flat gain response at low frequencies. Lowpass matching network topologies do not have any effect at low frequencies; therefore, we can use them to boost the gain where negative feedback does not help any more.

Another important issue is proper modeling of the passive components. In our previous examples, we were able to feed dc bias through reactive matching elements and the bias circuitry was decoupled from the RF signal path. In broadband feedback amplifiers the feedback and bias circuitry must also be included in the RF simulation.

2.8.4.1 Feedback amplifier design procedure

The major steps of broadband feedback amplifier design are as follows:

1. Examine the gain-phase characteristics of the active device. The frequency where the open-loop s_{21} phase crosses 90° should not be less than 40% to 50% (which are empirical limits) of the desired upper corner frequency of the feedback amplifier.
2. Select and compute series and/or parallel feedback resistors using (2.28) and (2.29). Add the primary parasitics (series L or parallel C) to the resistors to create RF component models.
3. Combine the feedback and dc bias circuits
4. Simulate and optimize the amplifier for the desired goals. Feedback is very effective at the lower RF frequencies where we have lots of reserve gain. If the goals are not reached, check the input and output ports to see if matching is needed.
5. If impedance matching is required, find the appropriate lowpass type circuit(s) to apply simultaneous conjugate match.
6. Reoptimize the overall circuit.

2.8.4.2 Feedback amplifier design formulas

Since negative feedback can only reduce the basic $|s_{21}|$ of the active device, we need to select a transistor that has a sufficient 50-Ω gain to meet our goal. Bipolar transistors have high transconductance, and they can easily provide 20- to 30-dB gain in the 50-Ω system. Although there are exceptions, FETs generally have lower basic gain, and 6 to 8 dB is a more typical range for them.

Approximate resistor values for a desired decibel gain of $G_{\mathrm{dB}AMP}$ between Z_0 terminations, using both series and parallel feedback of Figure 2.45, can be computed using

$$R_p \approx Z_0\left(1 + G_{AMP}\right) \tag{2.28}$$

$$R_S \approx \frac{Z_0^2}{R_p} - \frac{2Z_0}{\left|s_{21\mathrm{L}}\right|} \tag{2.29}$$

where

$$G_{AMP} = 10^{\frac{G_{\mathrm{dB}AMP}}{20}} \tag{2.30}$$

and $|s_{21L}|$ is the magnitude of the transistor's basic gain, at the lowest applicable frequency of the amplifier.

2.8.4.3 Illustrative example: design of a 10- to 4,000-MHz feedback amplifier

We want a cascadeable (well-matched) amplifier with 10-dB broadband gain to operate from 10 to 4,000 MHz in a 50-Ω system. Gain flatness of the circuit with nominal design values is to be within ±0.25 dB. Design goals for input and output return losses are to be better than 20 dB.

The Infineon BFP 520 is a good candidate for the task because it has $|s_{21}| = 31.6$ or 30 dB at 10 MHz (Figure 2.43), dropping to 5.76 or 15.2 dB at 4 GHz. Excluding dissipative losses in the feedback elements, with this device we can have feedback loop gain of 20 dB at the low end and more than 5 dB at the high end of the frequency range. The 90°, s_{21} phase crossing takes place at 1.82 GHz, but it will move higher when both feedback elements are added. The transistor was characterized at 2V, 20-mA dc bias, and we have a 9-V power supply available for the amplifier.

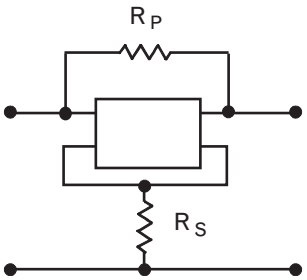


FIGURE 2.45 *Dual feedback applied to an active device can control gain as well as the input and output impedance. Since the feedback resistors are in the RF signal path, their parasitics must also be included during simulation.*

Computing the feedback resistors from (2.28) through (2.30),

$$G_{AMP} = 10^{\frac{G_{dBAMP}}{20}} = 10^{\frac{10}{20}} = 3.16$$

$$R_p \approx Z_0\left(1 + G_{AMP}\right) = 50\left(1 + 3.16\right) = 208\Omega$$

$$R_S \approx \frac{Z_0^2}{R_p} - \frac{2Z_0}{\left|s_{21L}\right|} = \frac{50^2}{208} - \frac{2\left(50\right)}{31.6} = 8.85\Omega$$

Using the nonlinear SPICE model of the transistor [27] and curve-tracer simulation, we determined the base-emitter voltage and base current for 20-mA collector current as

$$I_B = 0.163 \text{ mA}$$

$$V_{BE} = 931 \text{ mV}$$

Now we can compute the three dc bias resistors from the dc bias circuit shown in Figure 2.46(a). Feeding 10% of the 20-mA collector current through the resistive base voltage divider, the voltage and current values of the bias circuit are shown in Figure 2.46(b).

$$R_E = R_S = 8.85\Omega$$

$$V_E = 20.16 \text{ mA}\left(8.85\Omega\right) = 178 \text{ mV}$$

$$R_3 = \frac{6{,}820 \text{ mV}}{22.16 \text{ mA}} = 308\Omega \Rightarrow 309\Omega$$

$$R_2 = \frac{1{,}070 \text{ mV}}{2.16 \text{ mA}} = 495\Omega \Rightarrow 495\Omega$$

$$R_1 = \frac{1{,}110 \text{ mV}}{2 \text{ mA}} = 555\Omega \Rightarrow 555\Omega$$

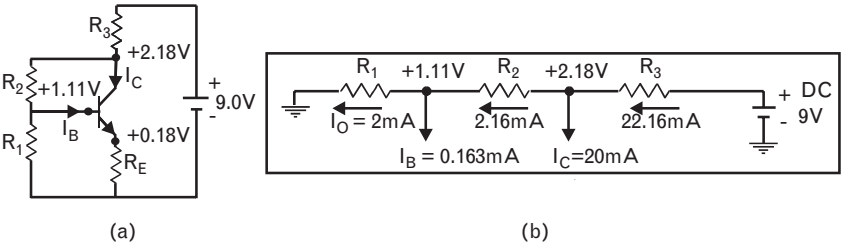


FIGURE 2.46 *(a) Resistive dc bias network with emitter and collector-to-base feedback assures high degree of bias stability. (b) Voltage and current distributions are rounded for the resistor calculations.*

Since the feedback resistors are part of the RF circuit, we need to use the actual physical models in the RF simulation. In our example, we estimated the minimum applicable series inductance for R_s as 0.2 nH, and 0.05 pF as the minimum parallel capacitance of R_p, since those are the dominant parasitic elements.[13] Figure 2.47 shows the combined RF and dc circuitry with their dominant parasitic reactances. The 495-Ω dc bias resistor, R_2, is split into two parts since it is larger than what is required for the RF feedback: $R_{2\,dc}$ 287Ω is bypassed for RF with a large parallel capacitor, and the remaining $R_p = 208\Omega$ is left in the RF path. The circuit is now ready for the initial RF simulation.

Figure 2.48 shows that the simulated performance of the feedback section comes very close to our targets. Although the input and output impedances are lower than 50Ω, the return losses are better than the minimum specified 20 dB. However, we only have 9-dB gain instead of 10 dB. Actually, if the shunting effects of R_1 and R_3 could be excluded, we would most likely meet both the gain and impedance specifications.

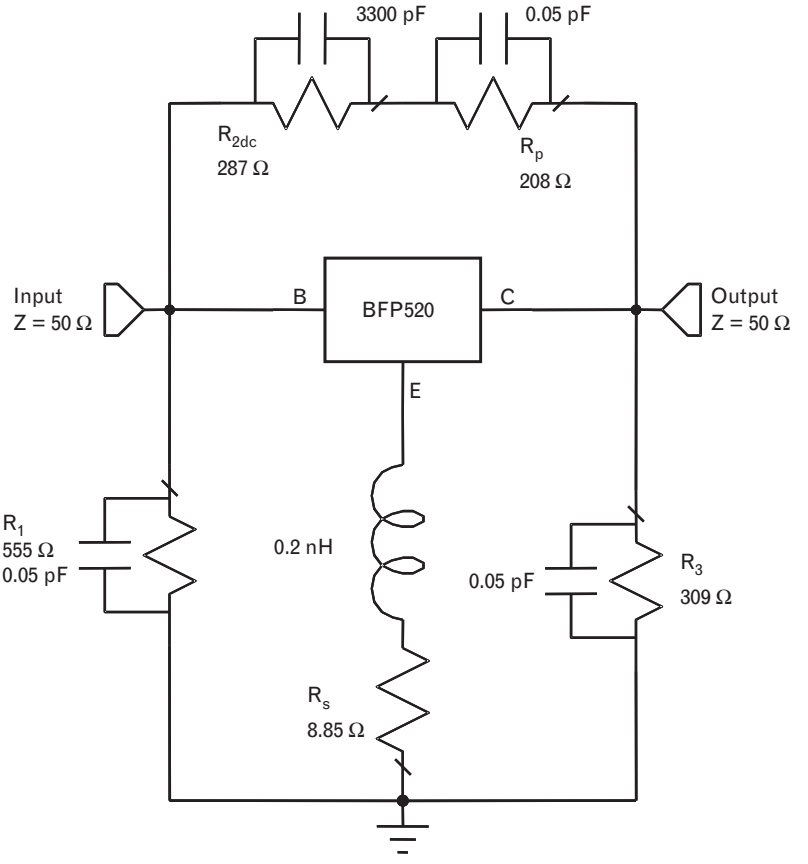


FIGURE 2.47 *RF equivalent circuit of the feedback amplifier. Elements ParFB and SerFB are the two RF feedback resistors with their parasitics. Elements R1 and R3 are part of the dc bias circuit. The 495-Ω collector-base dc resistor is shown in two parts: 208Ω (ParFB) is also used for RF feedback, but the remaining 287Ω (R2DC) is bypassed by a large value capacitor.*

13. RF component models are covered in Volume I, Chapter 7.

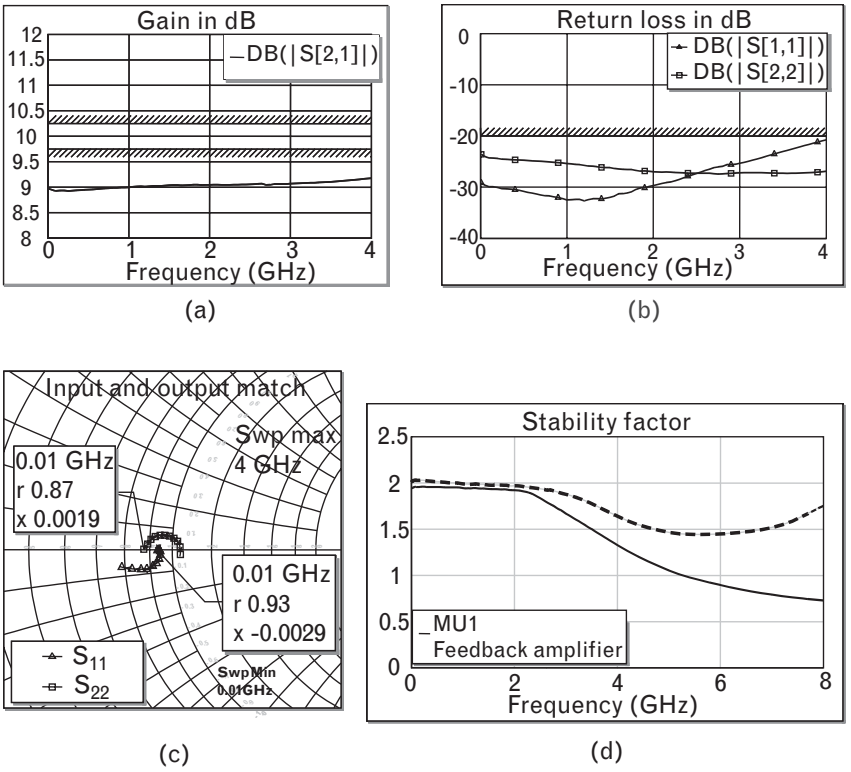


FIGURE 2.48 *(a) To meet our 10 ±0.25 dB gain we need to reduce the feedback, and that also affects the impedances. The positive gain slope may be a warning sign of RF stability problems. (b, c) Although we meet the return loss specifications, the Smith chart displays of s_{11} and s_{22} indicate that the low-frequency impedances are less than 50Ω. (d) The μ-factor drops below unity at 5 GHz, indicating stability problems. The second (dashed curve) response is showing the effect of added stabilization, as explained in the text.*

If we did not have the stability problem, we could increase the gain by reducing the amount of feedback. There are two possible adjustments of the feedback circuits to get more gain:

1. Decrease the series feedback resistor that also *decreases* the input and output impedances.
2. Increase the parallel feedback resistor that also *increases* the impedances.

Of the two choices, the second one adjusts the gain and impedances into the right direction. Increasing the parallel feedback resistor to 250Ω and slightly decreasing the series feedback resistor help to meet the specifications. (Manual tweaking may be lengthy and at this point our most effective option is circuit optimization.) However, the high-frequency stability

problem would not be solved. We either have to fix the excess phase-shift of the feedback loop or reduce the loop gain.

One relatively simple solution is to add some loss into the feedback loop and lower our gain specifications. Both input and output impedances are less than 50Ω; therefore, we need to add series, rather than parallel, resistance. Since the output impedance is the lower of the two, we add more resistance to that side.

Adding $R_{STAB1} = 4.7\text{-}\Omega$ series resistance to the input and $R_{STAB2} = 10\Omega$ to the output helps to balance the two impedances. After lowering our gain specifications to 9 dB, we submit the circuit to optimization. After a promising initial run, we add the dominant component parasitics and run a final optimization.

The complete amplifier circuitry after optimization (Figure 2.49) includes input and output coupling capacitors with their self-inductances. The 3,300-pF capacitors for coupling and decoupling represents $-j5\Omega$ reactance at 10 MHz, limiting the low end of the frequency response. Increasing capacitance extends the bandwidth but makes it more lossy in the gigahertz region. The 0.3-nH self-inductances of the capacitors are also included for more accurate simulation.

The amplifier is biased from a single 9-V dc supply. It is not suitable for portable wireless applications because a significant portion of the dc power consumption ($I_C^2 R_C = 150$ mW) is dissipated in the 309-Ω collector resistor. For the specified collector current, the dissipation can only be reduced by using smaller collector resistance. Unfortunately, this collector resistor shunts the RF output of the amplifier and cannot be reduced without lowering the output impedance of the amplifier.

Simulated results of the final amplifier (Figure 2.50) clearly display the benefits of negative feedback. Gain flatness and impedance match are maintained through more than two decades of frequency. To reassure ourselves about stability, we performed a small-signal Nyquist test that also confirmed stable operation. Although the closed-loop's gain function encircles the –1 point of the complex polar plane, the path from negative to positive frequencies follows a counterclockwise direction. The μ-factor of the amplifier, checked up to 8 GHz, confirms unconditional stability.

Noise performance of a resistive type of feedback amplifier is significantly degraded by (1) the resistors connected to the base and emitter of the device, and (2) the fact that the source termination of the device is not Γ_{OPT}. Computed noise figure of the broadband amplifier is 4.4 dB—nearly 3 dB higher than the optimum noise figure of the device.

We can apply negative feedback without such a high increase in noise figure by using broadband RF transformers instead of resistors [28, 29]. Additional benefits are less dc power dissipation and possible reduction of nonlinearities [30]. Disadvantages of the approach include difficulties of transformer modeling and production repeatability.

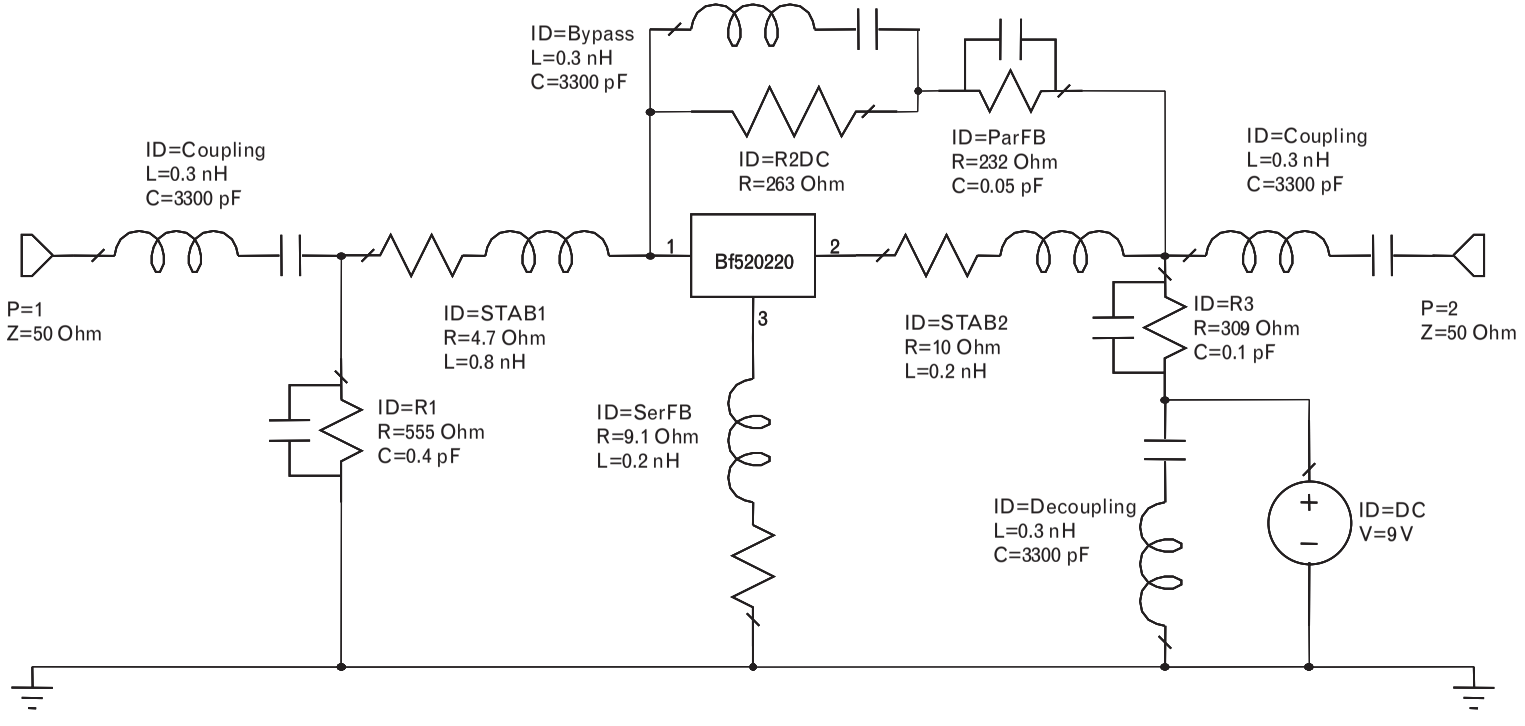


FIGURE 2.49 *Final RF schematics of the 10- to 4,000-MHz feedback amplifier, also showing input and output coupling capacitors and power supply bypass. The two resistors between base and collector provide dc feedback.*

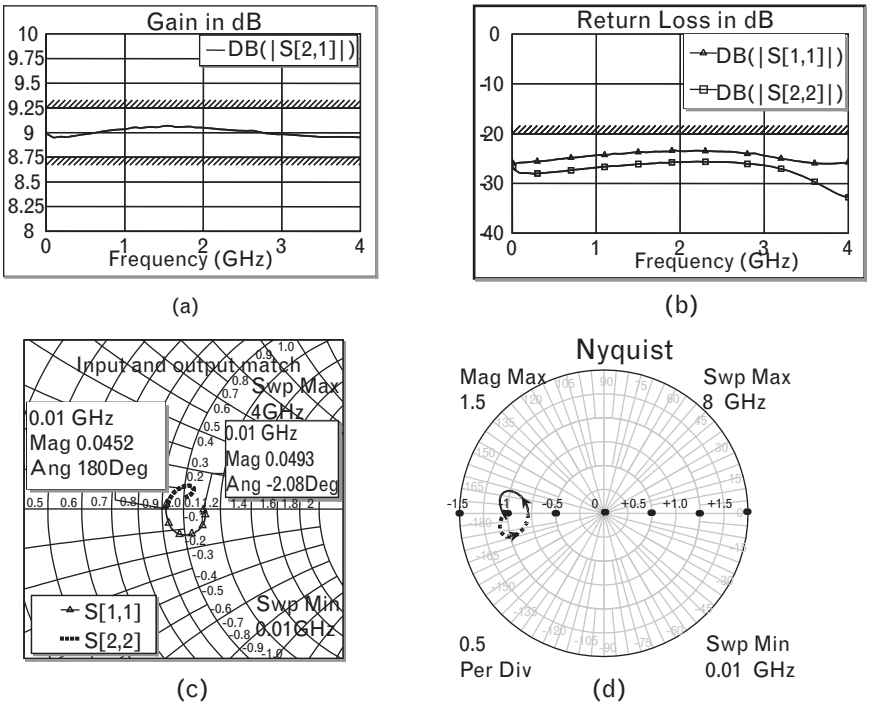


FIGURE 2.50
Final RF simulation of the feedback amplifier using simple physical component models. (a) Gain, (b) input and output return loss, and (c) s_{11} and s_{22} (dotted curve) on the expanded Smith chart. (d) Nyquist plot of the feedback loop encircles −1 in counter-clockwise direction (dotted curve represents negative frequencies), indicating stability.

2.8.4.4 Component tolerance effects

One of the benefits of negative feedback is the reduced circuit sensitivity to component value variations. In broadband amplifiers the effect is particularly noticeable at the lower frequencies where the feedback loop-gain is relatively high. To illustrate this effect, we performed a Monte Carlo analysis on the 10- to 4,000-MHz amplifier's gain, using the following tolerances:

- ±1% for all resistors;
- ±5% for all component parasitics;
- ±5% for all *S*-parameter magnitudes, except for s_{21}, where we applied ±20%;
- ±5° for all *S*-parameter phase angles.

Not having actual sample data for the components, we specified normal distributions with 3σ set to the above-listed tolerance limits, without any tolerance correlation.

Plotting the gain of the amplifier versus frequency (Figure 2.51) shows an extremely tight control below 500 MHz, but even at 4 GHz the total variation would easily meet a production specification of ±1 dB.

Feedback amplifiers can be very useful in broadband RF systems. Stability is a major concern since transistors provide gain to 20 to 30 GHz and positive feedback is likely to occur at some frequency in most cases. If an RF stability problem shows up, a thorough investigation, including a Nyquist test with nonlinear models, is highly recommended.

2.8.5 Distributed amplifiers

Increasing the bandwidth of a single amplifier stage inherently decreases its gain. For broadband applications the single-stage gain may be quite low. Cascading individual amplifiers to obtain high broadband gain is eventually limited by the basic gain-bandwidth product of the active devices and the cascading process. After four to five cascaded stages, additional stages bring very little, or no improvement at all in the achieved bandwidth. In contrast, the bandwidth of distributed amplifiers[14] [31, 32] improves with the increased number of stages by a wide margin [33].

In a distributed amplifier, we form two transmission lines by using external series inductors with the parallel parasitic gate and drain capacitances of the active devices (Figure 2.52). FETs are more suitable for this type of application because their input and output ports have higher Q's

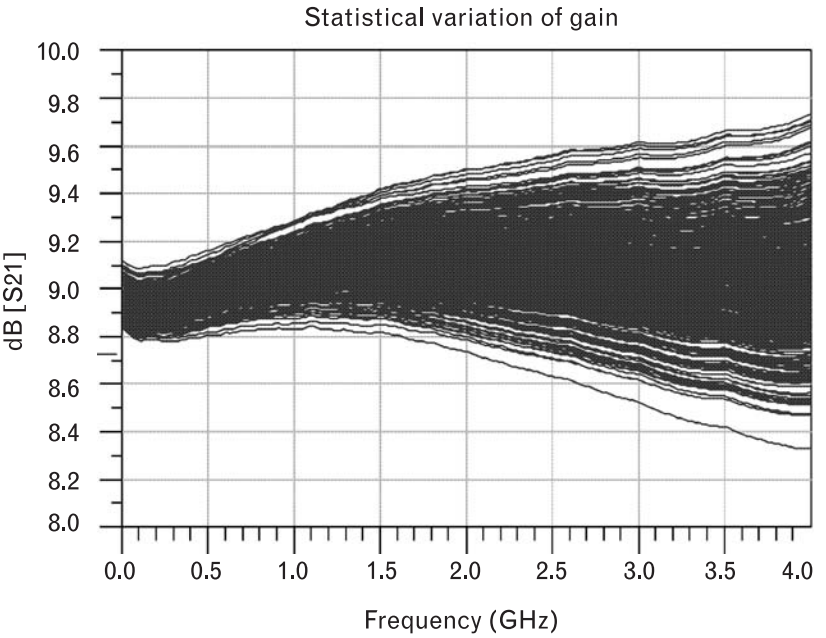


FIGURE 2.51
Monte Carlo statistical analysis with ADS can predict production line performance. Although we varied s_{21} of the transistor by ±20% at all frequencies, the amplifier's gain changes are much less, particularly at low frequencies where the device has higher open-loop gain.

14. Also called traveling-wave amplifiers, not to be confused with TWT, which is traveling-wave tube.

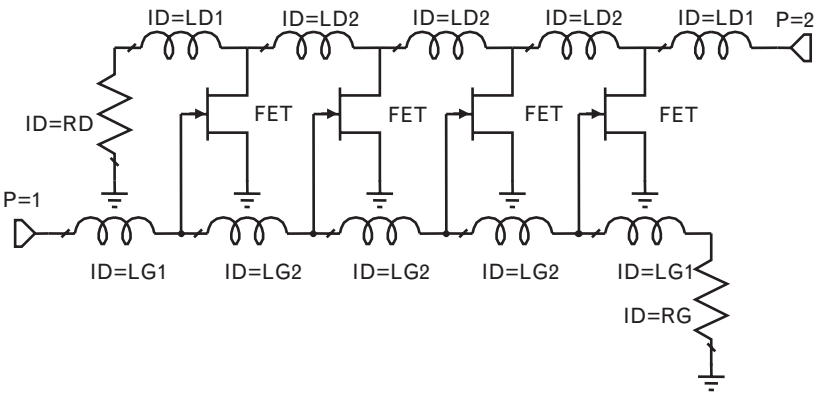


FIGURE 2.52 *In distributed amplifiers the input and output capacitances of the transistors are used as the distributed incremental parallel capacitors of two artificial transmission lines. Two sets of uniform series inductors (L_G and L_D) among those capacitors serve as the incremental inductances of the transmission lines.*

than bipolar transistors. The RF signal is coupled and amplified from one line to the other with the controlled sources of the devices. The input signal is distributed along the gate line and the amplified output is collected along the drain line. Proper terminations of these artificial transmission lines enable us to obtain much higher bandwidths than with comparable cascaded amplifiers. Although the concept was introduced in the 1930s and first used with electron tubes, it is now mainly applied in *RF integrated circuits* (RFICs), and we refer to the literature specializing in that technology [34–36].

If the input and output ports of the FETs used in the distributed amplifier were lossless, the number of cascaded stages and correspondingly the bandwidth would be unlimited. With real physical devices, there are practical limits. For high gain requirements, it may also be more economical to cascade a number of distributed sections [35]. For example, if a four-section distributed amplifier has a gain factor of 20, doubling the number of sections doubles the gain to 40. If instead we cascade two distributed amplifiers, the gain increases to 400, with a small reduction of bandwidth. Expecting the cascaded bandwidth reduction, if the distributed amplifier were initially designed for more bandwidth with somewhat less gain, the two cascaded five-section distributed amplifiers can still outperform the single 10-section unit.

2.9 Summary

In this chapter we focused on various types of amplifiers where the linear, small-signal *S*-parameters accurately characterize the active devices. We

need to realize, however, that *S*-parameters are not just frequency dependent, but they also vary with temperature and bias changes. In the rest of the book we go beyond linear operation and examine amplifiers operating in nonlinear modes. Since the *S*-parameters are not sufficient under those conditions, we also need new models to characterize active devices—topics we cover in our next chapter.

2.10 Problems

For all listed problems, download from http://www.infineon.com the broadband two-port *S*-parameters and noise-parameters of the Infineon BFP 405 transistor at 2-V, 2-mA bias condition. Use any available RF circuit simulator to perform the calculations. The AppCAD program is available through http://www.agilent.com.

1. Using the stabilized BFP 405 of Problem 2 in Chapter 1, design an amplifier stage for G_{MAX} at 880 MHz with ideal lumped elements. How does the gain compare to what you obtained in Chapter 1, Problem 1? Are the input and output VSWRs less than 1.5 for the 815- to 960-MHz band? If, not, increase the order of matching sections until you can meet those specifications.

2. Redesign the amplifier of Problem 1 for minimum noise. Check the noise performance of the low noise amplifier versus your results from the maximum gain amplifier. What is the magnitude of s_{11}? Is it possible to lower $|s_{11}|$ without a *significant increase* of the noise figure?

3. Design a balanced amplifier stage using two of the low noise amplifiers obtained in Problem 2. Download the Anaren 1E1304-3 (or any other similar) three-port hybrid coupler data from http://www.Anaren.com. How does the input and output reflection coefficient of the balance amplifier compare to the single stage of Problem 2? What kind of gain and noise sacrifice do you see by adding the directional couplers?

4. Apply negative feedback to the BFP 405 and design a 12-dB gain lumped element amplifier stage with ±0.5-dB gain flatness, between 10 and 2,000 MHz. What are the best input and output reflection coefficients you can get for this bandwidth? Design a dc bias network using a 9-V battery, operating the device at $V_{CE} = 2$V and $I_C = 2$ mA. Cascade two stages and observe the change of the gain flatness. Does the two-stage amplifier also meet the ±0.5-dB gain flatness specification?

References

[1] Application Note 154, "*S*-Parameter Design," Hewlett-Packard, April 1972.

[2] Gonzalez, G., *Microwave Transistor Amplifiers Analysis and Design,* 2nd ed., Englewood Cliffs, NJ: Prentice Hall, 1997.

[3] Gonzalez, G., *Microwave Transistor Amplifiers Analysis and Design,* 2nd ed., Englewood Cliffs, NJ: Prentice Hall, 1997, Appendix F.

[4] Gonzalez, G., *Microwave Transistor Amplifiers Analysis and Design,* 2nd ed., Englewood Cliffs, NJ: Prentice Hall, 1997, Appendices K and L.

[5] Ha, T. T., *Solid State Microwave Amplifier Design*, New York: Wiley-Interscience, 1981.

[6] Cripps, S. C., *RF Power Amplifiers for Wireless Communications*, Norwood, MA: Artech House, 1999.

[7] Stancliff, R., and D. Poulin, "Harmonic Load Pull," *IEEE MTT Int. Micr. Symp.Digest*, 1979.

[8] ATN Microwave, "A Load Pull System with Harmonic Tuning," *Microwave Journal*, March 1996.

[9] Losee, F., *RF Systems, Components, and Circuits Handbook*, Norwood, MA: Artech House, 1997.

[10] Friis, H. T., "Noise Figures in Radio Receivers," *Proc. of IRE*, July 1944.

[11] Bennett, W. R., *Electrical Noise*, New York: McGraw-Hill, 1960.

[12] Abrie, P. L. D., *Design of RF and Microwave Amplifiers and Oscillators*, Norwood, MA: Artech House, 2000.

[13] Fukui, H., *Low-Noise Microwave Transistors and Amplifiers*, New York: IEEE Press, 1981.

[14] Besser, L., "Stability Considerations of Low-Noise Transistor Amplifiers with Simultaneous Noise and Power Match," *IEEE MTT International Microwave Symposium Digest*, 1975.

[15] Vendelin, G., "Feedback Effects on Noise Performance of GaAs MESFETs," *IEEE MTT International Microwave Symposium Digest*, 1975.

[16] Suter, W. A., "Feedback and Parasitic Effects on Noise," *Microwave Journal*, February 1983.

[17] Kurokawa, K., "Design Theory of Balanced Amplifiers," *Bell System Technical Journal*, Vol. 44, No. 10, October 1965.

[18] Lange, J., "Integrated Stripline Quadrature Hybrids," *IEEE Trans. on Microwave Theory and Techniques*, December 1969.

[19] Piper, I., et al., "Balanced LNA Suits Cellular Base Station," *Microwaves & RF*, April 2002.

[20] Application Note 1281, "A High IIP3 Balanced Low Noise Amplifier for Cellular Base Station Applications," Agilent Technologies, February 2002.

[21] Grebene, A. B., *Bipolar and MOS Analog Integrated Circuit Design,* New York: Wiley, 1984.

[22] Howe, Jr., H., *Stripline Circuit Design*, Dedham, MA: Artech House, 1974.

[23] Application Note 154, "*S*-Parameter Design," Hewlett-Packard, April 1972.

[24] Giacoletto, G. S., *Electronics Designers' Handbook*, 2nd ed., New York: McGraw-Hill, 1977.

[25] Kenington, P. B., *High-Linearity RF Amplifier Design*, Norwood, MA: Artech House, 2000.

[26] Maclean, D. J. H., *Broadband Feedback Amplifiers*, New York: Research Studies Press/John Wiley & Sons, 1982.

[27] BFP 520—NPN Silicon RF Transistor Data Sheet, Infineon Technologies, June 2000.

[28] Norton, D., and A. F. Podell, "Transistor Amplifier with Impedance Matching Transformer," U.S. Patent 3,891,934, Anzac Corp., June 1975.

[29] Mead, H. B., and G. R. Callaway, "Broadband Amplifier," U.S. Patent 4,042,887, Q-Bit Corp., August 1977.

[30] Norton, D., "High Dynamic Range Amplifier," U.S. Patent 3,624,536, Anzac Corp., November 1971.

[31] Percival, A. F., "Thermionic Valve Circuits," U.K. Patent 460562, July 1936.

[32] Pierce, J. R., and L. M. Field, "Traveling-Wave Tubes," and "Theory of Beam-Type Traveling-Wave Tube," *Proc. IRE*, June 1947.

[33] Ginzton, E. L., et al., "Distributed Amplification," *Proc. of IRE*, August 1948.

[34] Soares, R., (ed.), *GaAs MESFET Circuit Design*, Norwood, MA: Artech House, 1988.

[35] Wong, T. T. Y., *Fundamentals of Distributed Amplification*, Norwood, MA: Artech House, 1993.

[36] Pengelly, R. S., *Microwave Field-Effect Transistors*, Atlanta, GA: Noble Publishing, 1994.

Selected bibliography

Cripps, S. C., *Advanced Techniques in RF Power Amplifier Design*, Norwood, MA: Artech House, 2002.

Ludwig, R., and P. Bretchko, *RF Circuit Design, Theory, and Applications*, Englewood Cliffs, NJ: Prentice Hall, 2000.

Rohde, U. L., and D. P. Newkirk, *RF/Microwave Circuit Design for Wireless Applications*, New York: Wiley-Interscience, 2000.

Vendelin, G. D., A. M. Pavio, and U. L. Rhode, *Microwave Circuit Design Using Linear and Nonlinear Techniques*, 2nd ed., New York: John Wiley & Sons, 2003.

CHAPTER 3

Active RF devices and their modeling

In the analysis we have seen so far, it has suited us to model active devices as "black boxes," in which the devices are modeled simply by their observed terminal characteristics. In doing so, we have implicitly assumed that the devices respond in actual circuits the same way as they do in the measurement system. In other words, if the devices are measured in a 50-Ω system, with 50-Ω terminating impedances at the source and load, then we assume the device itself is unaffected by changes in the input or output voltages or currents that may result when we use other values of impedance at the terminals. This assumption underpins the entire basis of linear design when we represent active devices by a matrix formulation such as *Y*- or *S*-parameters. We have implicitly assumed that the device is independent of the circuit in which we embed it.

Ultimately, of course, this assumption breaks down under conditions in which the signal swing of the voltage or current at given terminals becomes excessive. As we transition from a small-signal to a large-signal regime, the way in which we model the device needs to change as well. Because the large-signal model needs to be consistent at small-signal levels as well, our approach will be to start with the large-signal models and work back towards self-consistent small-signal models.

Large-signal models are generally either based on the physics of the device, or on empirical measurements. Physics-based models make certain assumptions to obtain the differential transport equations that describe the current flow in the semiconductor. Such models scale well and help the device designer understand how to optimize a device for a given application. Only rarely, however, can they model all observed phenomena well. Empirical-based models, on the other hand, attempt to curve fit measured data with either polynomial or functional equations that express the observed relationship between the device current and voltage. These models, while often accurate, can require extensive measurement, and they require a parameter extraction procedure in order to fit the equations to the data.

We shall not attempt to duplicate the analyses of many very good textbooks on device modeling in this chapter. Rather, the intent is to give

sufficient understanding of the device and its technology so the reader can understand some of the considerations in selecting a device for a given application, and understand the models required to accurately simulate the device in a given component.

3.1 The diode model

Because diodes are almost second nature to most electrical engineers, we shall only give a cursory treatment of the diode model here, pointing out some of the peculiarities of their use at high frequencies.

Standard semiconductor textbooks contain full treatments of the common p-n junction, the basis of most semiconductor diodes. The low-frequency diode equation

$$I = I_S\left[\exp\left(\frac{qV_J}{nkT}\right) - 1\right] \tag{3.1}$$

will be familiar to all electrical engineers, where the diode current I is a function of the junction voltage V_J across the diode and n is the ideality factor. For an ideal step-profile junction, $n = 1.0$, while for practical diodes it can be higher, perhaps up to 1.4. The k is Boltzmann's constant (1.37×10^{-23} J/K) and T the absolute temperature. I_s is a very small quantity; for instance, if the forward current is 1 mA at 0.3V at room temperature, then $I_S = 1 \times 10^{-8}$A. Such an equation models rectification behavior; at large negative voltages, the reverse current is I_S, and at forward voltages the current becomes exponentially large.

The p-n junction is a minority carrier device; when a forward voltage is applied, electrons are injected from the n region into the p region (and vice-versa with holes from the p region), where they are minority carriers. This excess density of minority carriers, or *stored-charge*, is modeled by a capacitance across the junction, and because minority carriers have low mobility and long transport times, it restricts the ultimate frequency at which the diode behaves as a rectifier. Because the carrier density is a function of the applied bias voltage, such devices can be used as variable capacitors, or varactors. In oscillators, for example, the ability to vary a reactance to tune to a particular frequency is particularly useful. Semiconductor diode junctions can be doped in such a way as to obtain a capacitance versus voltage profile that results in linear frequency tuning with voltage. If $C_J(0)$ is the capacitance of the an ideal p-n diode with zero applied bias, then the capacitance at other voltages is given by the familiar expression

$$C_J(V_J) = \frac{C_J(0)}{\sqrt{1 - \frac{V_J}{\phi}}} \tag{3.2}$$

where ϕ is the built-in potential difference across the diode. For silicon diodes it is typically 0.6V, and for GaAs diodes it is around 0.75V. For non-ideal diodes, the doping profile can be adjusted to give an exponent different from 1/2 in the denominator.

Similar equations can also be derived for the Schottky barrier diode, which is a much faster diode than a semiconductor junction. Here, a metal (such as Au, Pt, or Cu) takes the place of the *p*-material at the anode and is deposited onto an *n*-type epitaxial layer of either silicon or GaAs. The cathode connection is formed by bonding to a much more heavily doped layer of *n*+ semiconductor material in the substrate underneath the epitaxial layer. This second metal-semiconductor junction is not a rectifying junction because the built-in potential barrier is made negligible by the high n+ doping; it is known as an *ohmic* contact. The Schottky diode is a rectifying device up to very high frequencies because it is a majority carrier device; the conduction current consists entirely of free electrons, which are injected from the semiconductor into the metal.

The equivalent circuit of the diode is shown in Figure 3.1, where the parasitics from any package have been neglected. The conduction current is modeled as a current source according to (3.1), and the stored charge as a capacitance given by (3.2). The series resistance, typically modeled as a constant, results from the ohmic contact and semiconductor resistivity, which are both affected by the RF skin effect. The frequency limit of the diode is then inversely proportional to the product of R_S and C_J. Increasing the anode area or the doping density can decrease the resistance, but this will increase the capacitance. There is a trade-off, but in general the doping density is typically kept small to minimize the capacitance. This allows Schottky diodes to perform well into the high millimeter-wave region where other two-port devices have yet to penetrate. GaAs is the preferred material over silicon at microwave and millimeter-wave frequencies

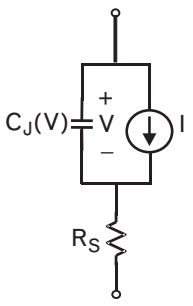


FIGURE 3.1
Equivalent circuit of a diode junction.

because of its higher electron mobility and better conversion performance in mixers.

3.2 Two-port device models

3.2.1 The output terminals of a two-port RF device

RF two-port devices have some surprisingly common characteristics at the most basic level. Figure 3.2(a) shows the output current of an ideal transistor plotted against the output voltage. For instance, for a GaAs MESFET, the plot shows the drain current as a function of the drain-source voltage.

This curve, known as the I-V curve of the device because it relates output current to output voltage, forms two distinct regions. The first region is at low values of output voltage, where the current increases linearly with output voltage; and the second region is at higher values of output voltage, where the current curves are flat with output voltage, or horizontal. The inflection point between the two regions is sometimes referred to as the *knee* of the curves.

Most amplifiers operate in the flat region of the I-V curve, and for now we will focus only on it. In this region, as the output voltage is increased, there is no change in the output current. This is typical of a current source: we may impose any voltage across a current source and the current is invariant. Its impedance (or output impedance, since we measure it at the output current terminals) is infinite. We see also that there is a family of

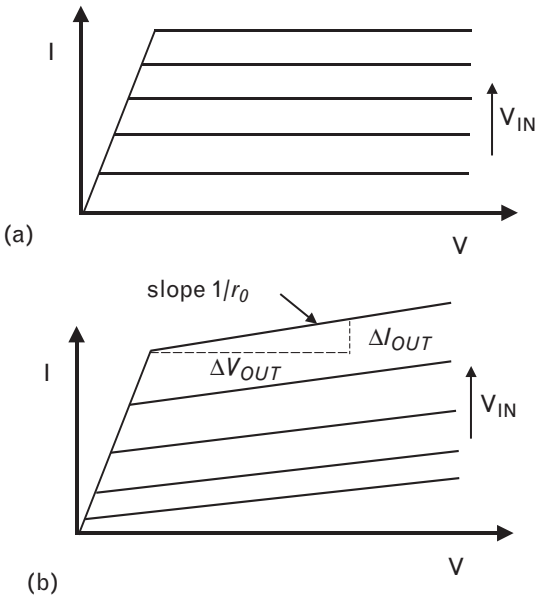


FIGURE 3.2
Curves of output current and voltage for (a) an ideal device and (b) a real device.

curves plotted in the flat region, each representing a particular value of control voltage at the input—for instance, the gate-source voltage for an FET. In the ideal case, as the control voltage is changed slightly, the output current changes in direct proportion to the input voltage change and the spacing between the horizontal lines of current is constant. If we increase the control voltage by 1V, the output current increases by the same number of milliamps each time, independent of the starting (or bias) point. If we define the transconductance g_m of a device as

$$g_m = \frac{\Delta I_{OUT}}{\Delta V_{IN}} = \frac{\partial i_O}{\partial v_{IN}} \tag{3.3}$$

then in the ideal case g_m is constant. The I-V curves are equally spaced if we plot them for equal increments of the input, control voltage.

We can also plot the *transfer characteristic* of the device to graphically show the relationship between the output current and input voltage at one, fixed, output voltage. Figure 3.3 shows such a plot for a common source MESFET. For the MESFET, V_{DS} is the output (drain-source) voltage and I_{DS} is the output drain-source current. V_{GS} is the control voltage, between the gate and source. As V_{GS} is reduced below zero towards the device pinch-off voltage $-V_p$, the drain current reduces from its maximum value I_{MAX} or I_{DSS} towards zero in direct proportion. As a result, its transfer characteristic on the left of Figure 3.3 is a straight line with slope g_m given by (3.3), which is a constant and in this case equals I_{MAX}/V_P.

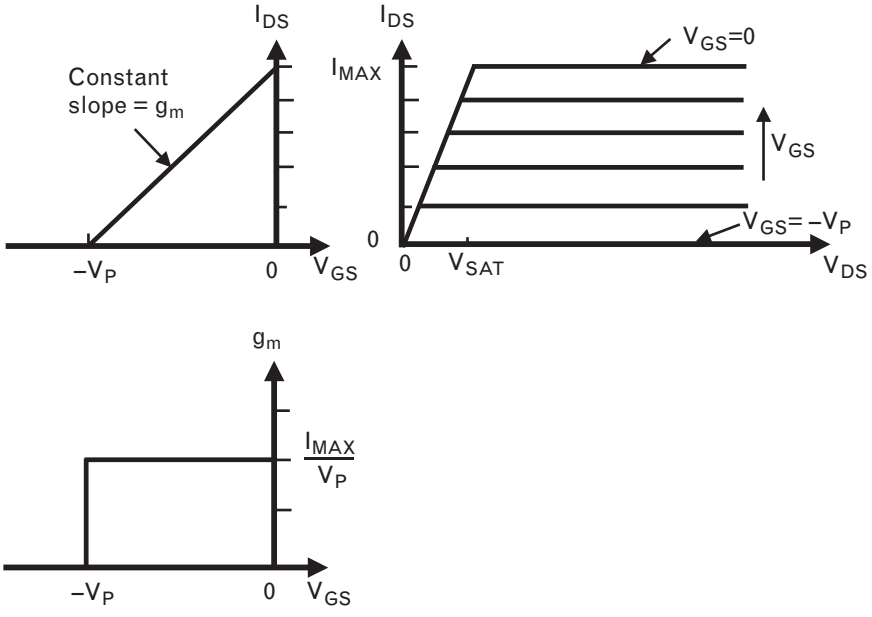


FIGURE 3.3
I-V curves for an ideal MESFET showing its transfer characteristic and transconductance.

The equivalent circuit model for device with the ideal characteristics of Figure 3.2(a) is shown in Figure 3.4(a). We know only that at the input we have a control voltage v_{IN} (where the lower case indicates we will consider incremental changes to current and voltage about the bias point), and that the output appears as a current source with incremental output current given from (3.3) by $g_m v_{IN}$. This current source is an ideal voltage-controlled current source where the transconductance g_m is a constant.

A more realistic device has I-V curves like those shown in Figure 3.2(b). Although the basic form is similar to an ideal device, the current lines are no longer horizontal, but dependent on the output voltage. If we consider the increment of current from the previous case (where the output current was flat) as ΔI_{OUT} at some value of output voltage, then we can define $\Delta I_{OUT} = \Delta V_{OUT} / r_o$, where V_{OUT} is the output voltage measured from the knee, and r_O is the reciprocal of the slope of the current curve. r_O is analogous to an output resistance, with ΔV_{OUT} across it and current ΔI_{OUT} through it.

Of course, not all devices are voltage controlled. It is frequently more convenient to model the output of a bipolar transistor as a current-controlled current source, in which the base current controls the level of the output current. In such a case, the current gain β or the forward transfer matrix element h_{FE} is defined as the ratio of the collector current to the base current. If the (output current) spacing between the I-V curves for equal increments of input base current is constant, then β is constant. For bipolar transistors, this is sometimes a more convenient representation because, as we will shortly see, g_m is not constant for a bipolar transistor.

Figure 3.2(b) shows a second characteristic of nonideal devices: nonconstant g_m. As the input voltage increases, the current increases more quickly for the same increment of input voltage, assuming the plot shows current curves plotted for equal increments of input voltage v_{IN}. Nonconstant g_m that varies with the amplitude of the input voltage causes nonlinear distortion and its resultant side effects, including intermodulation distortion, adjacent channel interference, and cross-modulation. Clearly, the best place to start to minimize these effects is to use a device with constant g_m.

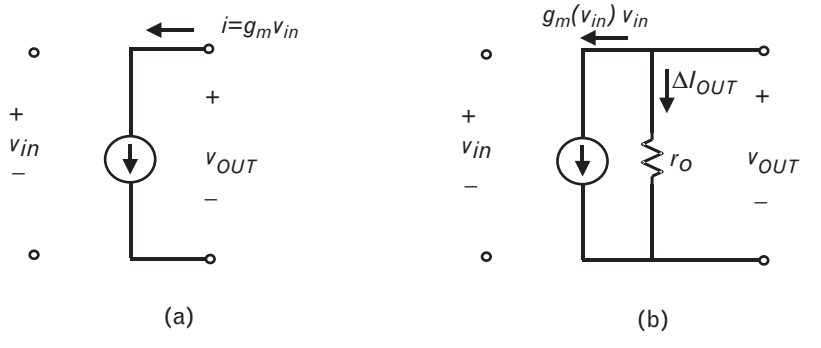


FIGURE 3.4
Equivalent circuit model fotr (a) the ideal device of Figure 3.2(a), and (b) he real device of Figure 3.2(b).

Figure 3.4(b) shows the equivalent circuit of the nonideal device. Now, the output current consists of two components: the current through the ideal current source, modeled as a voltage-controlled current source as before, and the component representing the incremental current ΔI_{OUT}, which flows through the output resistor r_O. Now, because g_m is not constant but a function of v_{IN}, the current source is a nonlinear component. Sometimes the slope of the I-V curves depends on the value of v_{IN} as well, and in such a case, r_O can also be a nonlinear function. In fact, in some models the nonlinear current source and output resistor are combined into a single current source of value i_O that is a function of *both* v_{IN} and v_{OUT}. In this way, the effective resistance of the current source looking into the output can be defined as

$$r_O = 1 / \frac{\partial i_O}{\partial v_O}\bigg|_{v_{IN}} \tag{3.4}$$

where these dependencies are now incorporated into the functional expression for the current. If the current source has no dependence on the output voltage, as in the ideal case, then the output resistance given by (3.4) is, of course, infinite.

As a final point, we must remember that the assumptions we have made above all relate to dc. In later sections, we will develop a fuller model for the output of the device based on the physics of the device itself, rather than on measured observation of its output terminal current and voltage at dc.

3.2.2 The bipolar transistor

At RF frequencies lower than 1 GHz, the silicon bipolar transistor was the first two-port device used for solid-state RF design. Because of the abundance of silicon-based processes in the digital world, and its lower thermal resistance and lower $1/f$ noise than GaAs, this device remains an important part of the designer's portfolio.

In deriving a model for the bipolar transistor based on the physics of the device, we will start with the simplest of models and gradually add to it in order to improve its accuracy. Throughout this section, we will use the terminology associated with a common-emitter NPN transistor for simplicity, and also because it is the most common type of RF transistor.

3.2.2.1 The Ebers-Moll model for the bipolar transistor

The best-known model of the bipolar transistor is the Ebers-Moll model, named after its original proponents. A cross-section of an NPN silicon bipolar transistor is shown in Figure 3.5.

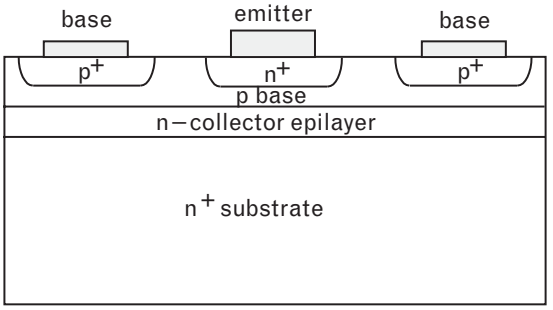


FIGURE 3.5
Simplified cross-section of an NPN bipolar transistor.

The collector-base and base-emitter junctions form p-n junctions and are modeled as simple diodes, while the external connections to the base, emitter, and collector are made to more highly doped regions of the semiconductor material and consequently form ohmic junctions. These ohmic junctions are essentially resistive in nature and allow the free flow of electrons and holes in either direction.

It would appear at first that the transistor is the back-to-back connection of two diodes. In fact, in the normal active region, the base-emitter junction is forward biased and the collector-base diode is reverse biased. In this way, the majority electrons from the *n*-doped emitter material are pulled into the *p*-type base region where they are now minority carriers. In normal diode operation, of course, these electrons would recombine with excess holes in the base and that would be the end of the story: Electrons would create a current as they flow from the emitter to the base. In fact, some electrons do recombine in the base, but because the base is thin, they come under the influence of the reverse-biased base-collector junction. The electrons are minority carriers in the base. In a reverse-biased diode there is normally no current because minority carriers are, in fact, rare, and those that do exist are quickly pulled to the opposite side of the junction. The collector is normally depleted of free carriers by the large applied electric field. The injected minority electrons are quickly swept from the base by the action of the high positive potential at the collector, which keeps the base-collector reverse biased. In this way, electrons flow from the emitter through the base into the collector, losing only a fraction due to recombination in the base region.

The current gain of the transistor, the ratio of collector current to base current, is ultimately limited because the base current is nonzero. This results from recombination of electrons in the base, and also from hole injection from the base itself into the emitter.

The applied voltages across each diode are shown in Figure 3.6. In the inverse region of operation, the roles of the collector and emitter are reversed from those in the active region, although with some loss of injection efficiency. Then, any electrons generated by the *n*-type collector that flow into the base are pulled into the emitter if the base-emitter junction is

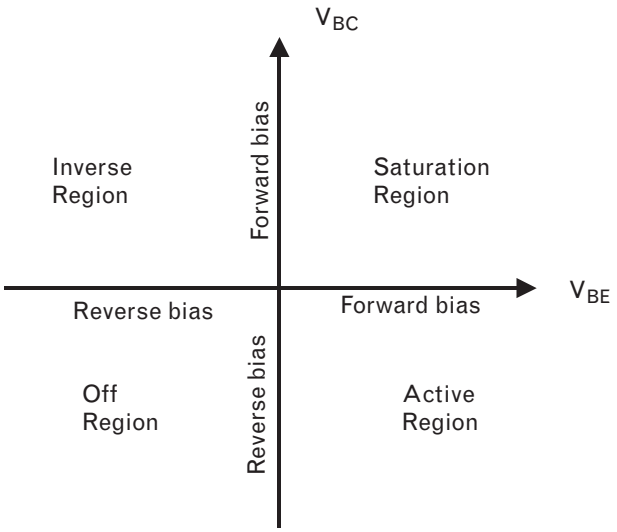


FIGURE 3.6
Applied voltages across the two diode junctions of the NPN transistor, showing its modes of operation.

properly biased. However, the reverse current gain is small because the collector doping is kept relatively low to minimize free charge (which minimizes capacitance and improves the speed), so not many electrons are injected into the base.

Such reverse injection is negligible in the normal active region, since the collector-base junction is reverse biased and any electrons generated in the collector are swept out of the collector terminal as majority carriers. However, more significant in the normal active region will be the impact of any holes generated in the collector, where they start life as minority carriers. This could constitute a significant reverse current flow because of the large potential between the collector and base. Normally, because the collector is doped with *n*-type material, this reverse hole current is very small since most of the holes generated in the collector will only arise from impurities at the collector-base interface, so the reverse current flow will only be due to leakage and thermal effects which are kept small. However, at elevated temperatures and as the collector-base junction approaches breakdown, this reverse current becomes very significant.

The transistor may therefore be modeled at dc by two diodes representing the p-n junctions themselves. We need to add two current sources as well, since the normal diode equation solves only for the current flow resulting from electrons or holes generated from within it; it cannot represent the additional current resulting from minority carrier injection. Thus, one current source I_{CC} represents the minority carrier current due to the electrons injected into the collector-base junction from the emitter; and the other I_{EC} represents the minority current that results from carriers injected from the collector. In the normal active region, the latter is very small since there are negligible free holes in the (*n*-type) collector and very few electrons can cross the large potential barrier created by the reverse bias

between the collector and base. Figure 3.7 shows the elements in the basic model.

The concept of recombination of electrons in the base is introduced into the model through the forward and reverse recombination factors α_F and α_R, respectively. The forward recombination factor is simply the ratio of the forward collector current I_{CC} to the forward emitter current, and is slightly less than one. Considering forward current only, the base current is simply the difference between the emitter current and the collector current, and it is due to the electrons that do not reach the collector from the emitter, but instead meet their match in the base. Recombination of an electron and hole in the base must create a terminal current in order to maintain charge neutrality; the resulting current flow is just the base current. The α_R is usually smaller than α_F, because the geometry of the device is normally optimized for the active, rather than the inverse region, and the emitter efficiency is optimized for electron injection rather than the collector efficiency.

The equations resulting from the model are then

$$I_{BE-diode} = I_{ES}\left(e^{\frac{qV_{BE}}{kT}} - 1\right) \triangleq \frac{I_{CC}}{\alpha_F} \tag{3.5}$$

$$I_{BC-diode} = I_{CS}\left(e^{\frac{qV_{BC}}{kT}} - 1\right) \triangleq \frac{I_{EC}}{\alpha_R} \tag{3.6}$$

Now semiconductor physics tells us that to maintain charge neutrality in the device when two semiconductor junctions are placed together, their Fermi energy levels must equalize. It can then be proven that

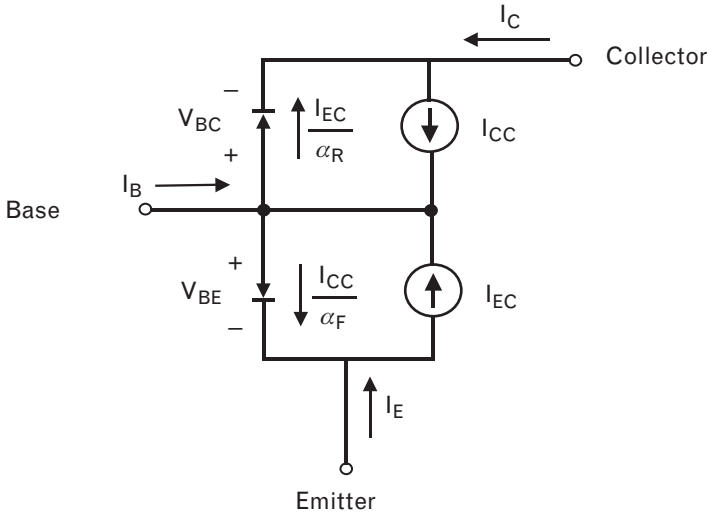


FIGURE 3.7
Basic dc model for an NPN bipolar transistor.

$$\alpha_F I_{ES} = \alpha_R I_{CS} \triangleq I_S \tag{3.7}$$

where I_S is known as the transistor saturation current. The resulting equations for the device become

$$I_{CC} = I_S \left(e^{\frac{qV_{BE}}{kT}} - 1 \right) \tag{3.8}$$

$$I_{EC} = I_S \left(e^{\frac{qV_{BC}}{kT}} - 1 \right) \tag{3.9}$$

The terminal currents flowing into the transistor can then be written:

$$I_C = I_{CC} + \left[-\frac{1}{\alpha_R} \right] I_{EC} \tag{3.10}$$

$$I_E = \left[-\frac{1}{\alpha_F} \right] I_{CC} + I_{EC} \tag{3.11}$$

$$I_B = \left[\frac{1}{\alpha_F} - 1 \right] I_{CC} + \left[\frac{1}{\alpha_R} - 1 \right] I_{EC} \tag{3.12}$$

These equations represent the dc currents that flow into the device, and can be calculated once the applied voltages are known. Represented in this form with $I_{EC} = 0$, the dc hybrid-T topology for the transistor can be readily derived.

We can jump ahead a little and foresee from this model how the small-signal T-model for the transistor in Figure 3.8 is derived. We can

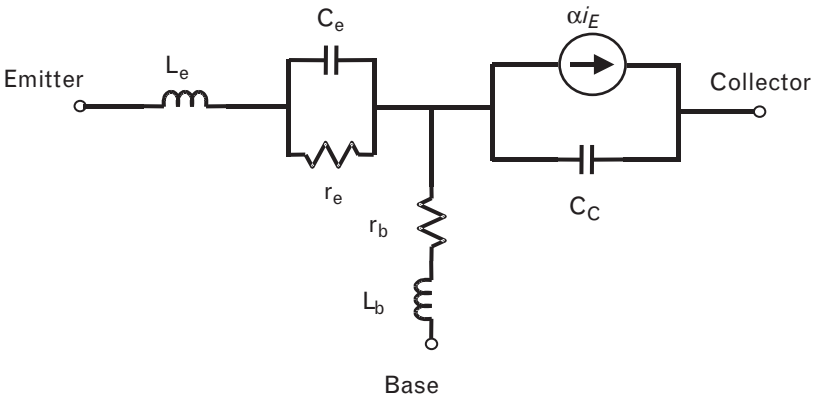


FIGURE 3.8
Basic T-topology model for an NPN bipolar transistor.

model the device at nonzero frequencies by accounting for the storage of minority carriers and representing them by capacitance. We can also model the current generator α as a function of frequency, to include two time delay terms to model the time an electron takes to transit through the base and the collector depletion region. The first, $\tau_b = 1/\omega_b$, represents the time that the minority carriers take to traverse the base. It dominates α and controls its 3-dB roll-off frequency (the base cutoff frequency). The second delay term τ_c represents the additional transit time through the collector depletion layer. τ_c increases with reverse collector bias as the collector depletion region thickness swells. Thus,

$$\alpha_F = \frac{\alpha_{F,dc} \exp(-j\omega\tau_c)}{1 + j(\omega/\omega_b)} \tag{3.13}$$

The transit times τ_b and τ_c are additive components to the total transit delay between the emitter and collector τ_{ec}. The other additive components to model the total transit delay τ_{ec} are the charging times of the base-emitter and base-collector capacitances through the emitter resistance; these are implicitly accounted for by the R-C topology of the model itself.

The T-model can prove useful in deembedding parasitic elements as part of the parameter extraction process used to determine the component values for more complex models. The T-model also proves useful in modeling *heterojunction bipolar transistor* (HBT) devices.

A simpler model results when the two current sources of Figure 3.7 are combined into a single current source between the emitter and collector as shown in Figure 3.9.

The two topologies in Figures 3.7 and 3.9 can be shown to be identical by equating the terminal base, collector, and emitter currents in each case.

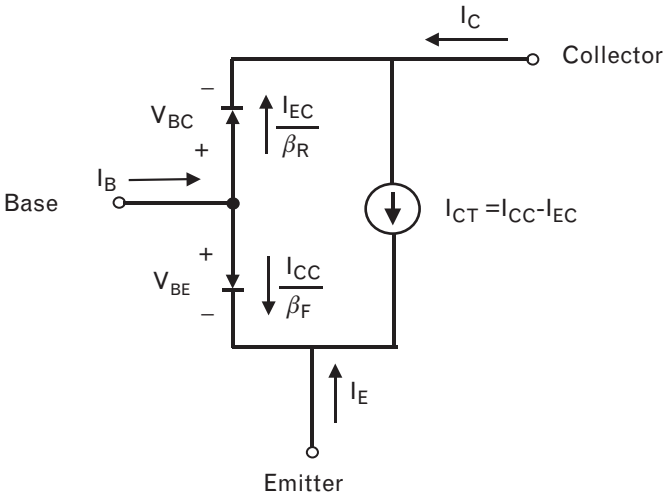


FIGURE 3.9
The Ebers-Moll circuit topology for a bipolar transistor.

The expressions for I_{CC} and I_{EC} are the same in both cases, but now the current source has a current I_{CT} defined by

$$\begin{aligned} I_{CT} &= I_{CC} - I_{EC} \\ &= I_S\left[\left(e^{\frac{qV_{BE}}{kT}} - 1\right) - \left(e^{\frac{qV_{BC}}{kT}} - 1\right)\right] \\ &= I_S e^{\frac{qV_{BE}}{kT}}\left[1 - e^{\frac{-qV_{CE}}{kT}}\right] \end{aligned} \tag{3.14}$$

in order to keep the terminal currents in the two topologies identical. This final expression establishes the typically observed exponential relationship between collector current and base-emitter voltage, and the "saturation" characteristic observed with the collector-emitter voltage V_{CE}. The currents that flow in the diodes are now given by I_{CC} / β_F and I_{EC} / β_R, where

$$\begin{aligned} \beta_F &\triangleq \frac{\alpha_F}{1-\alpha_F} \\ \beta_R &\triangleq \frac{\alpha_R}{1-\alpha_R} \end{aligned} \tag{3.15}$$

It can be seen from (3.10) and (3.14) that in the active region of the device, where V_{BC} is very large and negative, I_{EC} is vanishingly small, and from inspection of Figure 3.9

$$\begin{aligned} I_C &\approx I_{CC} \\ I_B &\approx \frac{I_{CC}}{\beta_F} \end{aligned} \tag{3.16}$$

so that β_F is just the ratio between the collector current and base current, or the common-emitter current gain of the device. The terminal currents now become

$$\begin{aligned} I_C &= \left(I_{CC} - I_{EC}\right) - \left[\frac{I_{EC}}{\beta_R}\right] \\ I_B &= \left[\frac{I_{CC}}{\beta_F}\right] + \left[\frac{I_{EC}}{\beta_R}\right] \\ I_E &= -\left[\frac{I_{CC}}{\beta_F}\right] - \left(I_{CC} - I_{EC}\right) \end{aligned} \tag{3.17}$$

We should return to Figure 3.6 for a moment to verify if the device model is applicable over the entire range of operating conditions. We have already discussed the conditions applicable in both the normal active region (fourth quadrant) and the inverse region (second quadrant) of the transistor. On the I-V curves of Figure 3.2(a), the normal active region corresponds to the section with the horizontal current curves (to the right of the knee); there, the output can be modeled by a current source with infinite output impedance. This region is referred to as the *linear region* of a bipolar transistor because in this region linear operation results, assuming the I-V curves are constantly spaced.

In the *saturation region* of Figure 3.6, both diodes are forward biased. This condition can apply momentarily in a linear amplifier when the collector voltage swings close to zero, as a result of the voltage drop from the bias rail induced by a large collector current flowing in the load. When the collector voltage swings close to zero under these conditions, the base-collector diode can become slightly forward biased, and electrons are therefore injected from the collector into the base (as well as from the emitter). As a result, the base becomes bloated with minority carriers (electrons) and the transistor is said to be saturated. Saturation corresponds to the region to the left of the knee in Figure 3.2. From (3.9), I_{EC} is no longer small, and the base current in (3.12) will have two significant components. Equation (3.17) shows that the (total) terminal collector current also starts to decrease under these conditions. Saturation is an important condition to recognize since it is highly nonlinear: the base becomes engorged with minority carriers, the stored charge (capacitance) increases rapidly, the transit times increase, and the output voltage becomes highly dependent on output current.

The knee voltage between the collector and emitter is therefore referred to as the saturation voltage V_{sat} of the device. It is an important parameter because it is the minimum voltage we can obtain between the collector and emitter in normal circuit operation, and limits the efficiency we can achieve from an amplifier. As can be seen from the figure, V_{sat} increases with collector current. Furthermore, it has a positive temperature coefficient, so that the device becomes even less efficient as the temperature increases.

The off region of Figure 3.6 (third quadrant) is when both base-emitter and base-collector diodes are reverse biased, and there is no injection from either emitter or collector.

Finally, we can include an additional equation from semiconductor physics to model the temperature dependence of the diodes:

$$I_S(T) = I_S(T_{nom})\left[\frac{T}{T_{nom}}\right]^3 \exp\left(-\frac{E_g}{k}\left(\frac{1}{T}-\frac{1}{T_{nom}}\right)\right) \tag{3.18}$$

where T_{nom} is the reference temperature (room temperature), and E_g the energy bandgap of the semiconductor.

This completes the Ebers-Moll equations. They form a basic model that is useful for dc characterization of a device. It is a simple model because the equations require only three model parameters, β_F, β_R, and I_S for complete characterization. Of course, E_g is also required to model the variation of I_S with temperature should this be required. Although an elegant starting point, unfortunately, the Ebers-Moll model neglects too many effects to be useful at RF frequencies.

3.2.2.2 Breakdown effects in the bipolar transistor

Real diodes, of course, eventually exceed the limits for which the simple diode equations of (3.5) and (3.6) are applicable, and ultimately fail. In principle, the Ebers-Moll equations could be modified to incorporate breakdown and other effects within the transistor model. In practice, it is the base-collector diode that designers need to be most aware of, since this diode is reverse biased and as the collector voltage increases, it ultimately reaches a threshold voltage V_{XYO} where avalanche multiplication begins and the reverse collector current component becomes dominant. In the notation most commonly used, the subscripts X and Y refer to the transistor terminals (base, emitter, collector) across which the voltage at collector-base junction breakdown is measured, and the O refers to the open-circuit condition of the remaining third terminal. Sometimes S is used instead of 0, indicating that the remaining terminal is short-circuited. Thus, V_{CBO} is the voltage between the collector and base at which breakdown of the collector-base junction occurs, when the emitter is open-circuited to current flow. The behavior of the collector current against V_{CB} is shown in Figure 3.10.

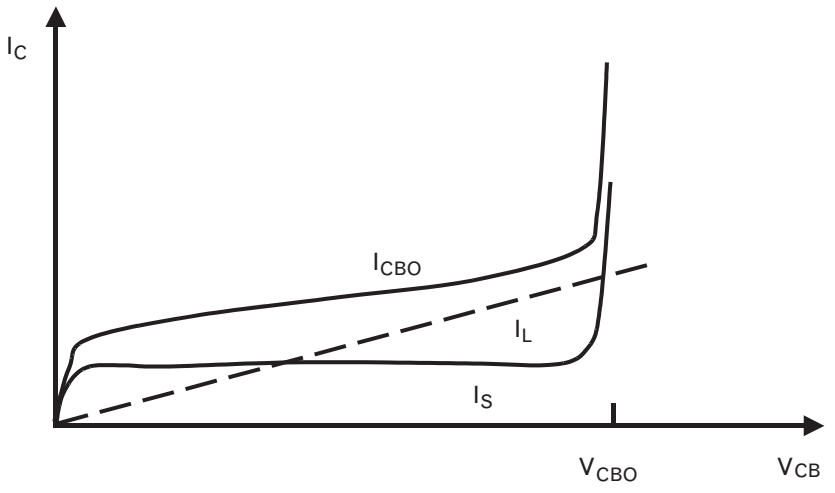


FIGURE 3.10
Reverse collector current of a bipolar transistor, showing the effects of the reverse bias collector voltage and breakdown.

V_{CBO} is typically similar to V_{CES}, the breakdown voltage measured between the collector and emitter when the base bias network has no resistance to the breakdown current (i.e., the base appears as a short circuit to the avalanche current from the collector). When the collector-base junction enters avalanche, the impedance between the collector and base is very low and large currents flow—typically leaving the transistor through the base terminal. Whether the emitter is open- or short-circuited has little impact on the breakdown voltage necessary for avalanche. If instead we measure V_{CEO}, however, the avalanche current cannot flow out the base but must flow through the transistor instead and out the emitter. In this case, the breakdown voltage V_{CEO} can be much less than V_{CES} or V_{CBO}, often less than half. In fact, $V_{CEO} \approx V_{CBO}(1-\alpha_F)^k$, where k is a positive fractional fitting factor less than one, so V_{CEO} decreases to zero as α_F increases toward unity (i.e., as β_F or the bias current I_C increases [1]). To minimize temperature effects on the bias current, most transistor bias circuits present a low internal resistance to the base, particularly at RF frequencies where any shunt capacitance will reduce the effective resistance seen by the base. The avalanche current can indeed freely flow out the base into the bias network, and an amplifier design can benefit from the higher breakdown voltage (V_{CBO}) that is then applicable. However, as the equivalent loading resistance on the base increases, the breakdown voltage will become lower, towards V_{CEO}, and the maximum power that the device can provide will become smaller. In practice, the applicable breakdown voltage will be some value intermediate between V_{CEO} and V_{CBO}, depending on the RF resistive load in the base terminal.

We see from Figure 3.10 that even before the onset of breakdown the leakage current, labeled I_{CBO}, can become appreciable, and especially with temperature. I_{CBO} can be a significant part of the base current (since the base is a low resistance path through which this leakage current can flow) and can contribute to self-heating and thermal runaway effects if not properly terminated, since if they cause the base voltage to increase, the effect is exacerbated. The reverse current consists of two components. The first, I_L, is a leakage component that results from surface traps (impurities) in the base-collector junction being pulled from the crystal lattice as the electric field increases with the reverse bias. These traps result in additional electron-hole pairs that then move across the junction and create a component in the reverse current. The second component, I_S, is due to the thermal generation of electron-hole pairs within the silicon crystal itself; while temperature dependent, it is not until the electric field reaches a threshold value that the electron-hole pairs achieve sufficient energy to begin knocking other pairs apart, and causing avalanche multiplication to begin.

Both breakdown and reverse leakage current need particular attention in the design of power amplifiers. Breakdown limits the maximum attainable collector voltage swing, and this is a fundamental limit of the

maximum output power that can be generated from a transistor. The reverse leakage current can cause the device to become hotter and reduce its efficiency, as well as result in destruction of the device due to thermal runaway.

3.2.2.3 Small-signal transistor model derived from the Ebers-Moll model

The topology of Figure 3.9 needs to apply equally well for small voltage variations about the dc bias point predicted by the Ebers-Moll equations above. A very useful approximation to the low-frequency behavior of the transistor can be derived by differentiating (3.14), and using (3.16) and (3.3) to obtain

$$g_m = \frac{q}{kT} I_S \left[e^{\frac{qV_{BE}}{kT}} \right] \approx \frac{q}{kT} I_C \tag{3.19}$$

In a small signal model, where all voltages and currents are modeled as incremental quantities about a quiescent value, the collector current source can be modeled using (3.3) as

$$i_C = \delta I_C = g_m \delta V_{IN} = g_m v_{BE} \tag{3.20}$$

In the active region, the forward-biased base-emitter diode can be modeled as a resistor r_π and the reverse-biased base-collector diode as a resistor r_μ. Because of the reverse-bias, r_μ is normally infinite. The r_π can be determined by applying a small incremental voltage δv_{BE} and calculating the incremental current δi_B that results. Using (3.16) we obtain

$$r_\pi = \frac{\delta v_{BE}}{\delta i_B} \approx \frac{\delta v_{BE}}{\delta i_C} \beta = \frac{\beta}{g_m} \tag{3.21}$$

This yields the small-signal incremental circuit of Figure 3.11. Because it is a dc model, it neglects any charge storage associated with the transit of the minority carriers through the base region. Later, this charge is modeled by capacitance. It also neglects any output resistance of the device (between the collector and emitter), since this is represented by a current source of infinite impedance. The next model for the bipolar transistor accounts for these and other effects.

3.2.2.4 The Gummel-Poon model

The Gummel-Poon model [2–4] improves on the Ebers-Moll model by considering a number of additional features:

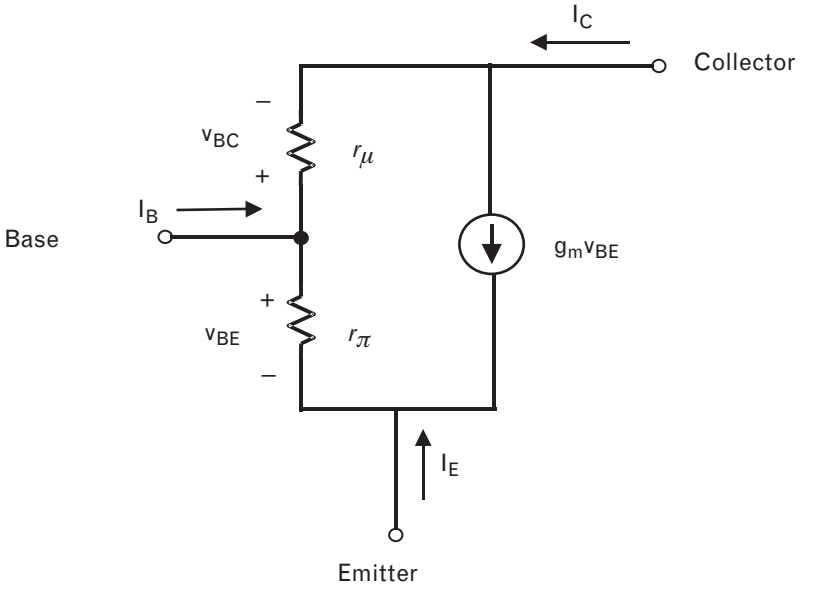


FIGURE 3.11
Incremental dc model for the transistor derived from the Ebers-Moll equations.

- *Low current effects*, which result in additional base current due to recombination of the minority carriers (electrons) in the base. This degrades the current gain.

- *High level injection*, which occurs when excess majority carriers (holes) spill over from the base into the collector because the retarding electric field at the base-collector junction vanishes at high currents. This is known as the Kirk effect. In essence, the concentration of injected minority carriers in the base now becomes significant compared with the majority carrier concentration, which itself must increase to maintain charge neutrality. The increase in collector current with input current begins to slow, and β is reduced.

- *Base-width modulation*, when the effective width of the base through which the electrons must transit is increased when the forward bias on either of the junctions increases. This results in an increase in the forward and reverse base-transit times. The base-collector capacitance is also increased because of the hole (majority carrier) charge in the extended base.

- *The ac model parameters*, to account for charge aggregation within the device, and for the intrinsic resistances associated with the ohmic junctions at the base, emitter, and collector, as well as resistance encountered by the minority carriers as they traverse the transistor. The latter is an intrinsic base resistance and is modeled as current dependent. The bias voltage is nonuniform across the base region and causes the electrons to aggregate in regions of higher electric field (current crowding). This causes hot spots within the device as well as variable

resistance. The capacitances are also functions of the applied voltage, thus nonlinear.

The Ebers-Moll model assumed a β_F (or β_R) that was constant. In fact, as stated above, the current gain is dependent on the total current that flows. If $V_{BC} = 0$ and the collector current I_C of (3.14) and the base current I_B of (3.5) are plotted against V_{BE}, an exponential relationship should result. Alternatively, a plot of the logarithm of these currents against V_{BE} should result in a straight line, with the difference between the two curves equal to $\ln(I_C) - \ln(I_B) = \ln(I_C / I_B) = \ln(\beta_F)$. Such a plot is known as a Gummel plot. In reality, Figure 3.12 represents the behavior that is modeled. It shows three distinct regions with representative values of voltage and (log) current indicated. It can be seen that the difference between the two curves is indeed dependent on the applied voltage.

Although there is a variation with voltage, the relationship of I_C to I_B (through β_F) is still more linear than that of I_C to V_{BE}. This is the reason that the I-V curves for a bipolar transistor are generally plotted with I_B as the controlling input parameter instead of V_{BE}, so that the difference between successive curves corresponds to the dc current gain β_F instead of g_m. The three regions are also shown in Figure 3.13, where β_F is explicitly plotted against collector current rather than input voltage. In the first region, β_F decreases at very low currents, because recombination of the minority carriers in the base and in surface channels adds additional components to the base current. As a result, its ratio with the output current is reduced. In the second region, both I_C and I_B depend exponentially on V_{BE} and the current

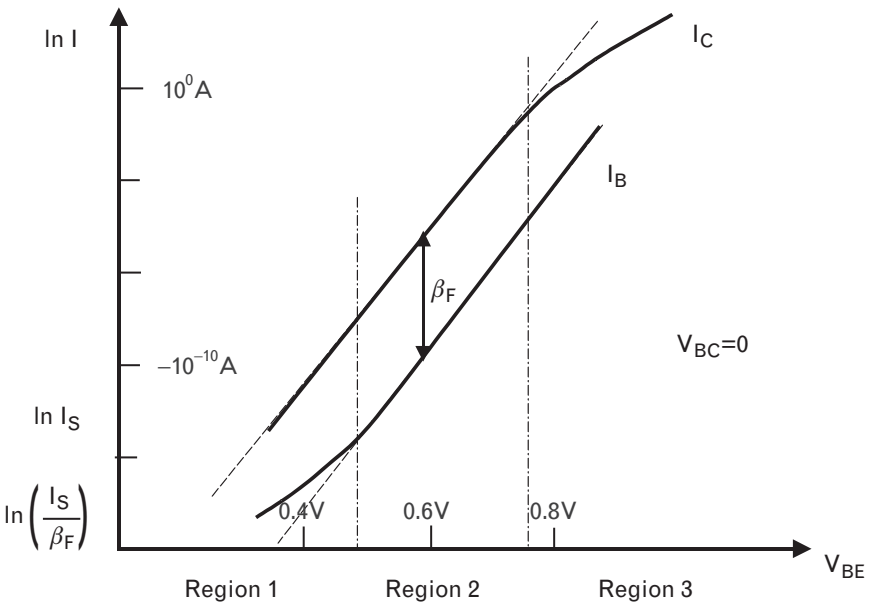


FIGURE 3.12
Variation of collector current and base current for a bipolar transistor as the base-emitter voltage is varied.

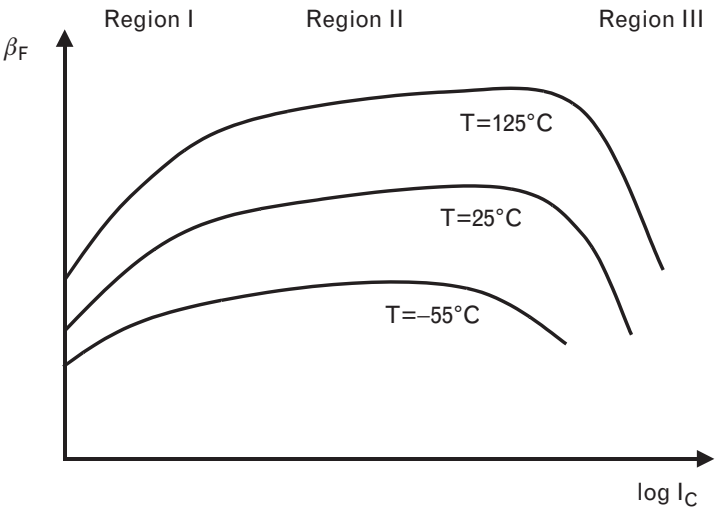


FIGURE 3.13
Variation of forward current gain with collector current.

gain is invariant to collector current. In the third region, I_C starts to decrease below the expected exponential behavior due to the effect of high-level injection, base-width modulation, and voltage drops across the base and emitter resistances, so the gain starts to decrease.

Another effect modeled by the Gummel-Poon model is the output resistance, which shows up as a nonzero slope on the output I-V curves. Effectively, as the reverse bias on the base-collector junction is increased, the depletion region (the region in the collector stripped of its electrons) becomes larger. This changes the effective width of the normally thin base and manifests itself as a change in the amount of incremental output current that flows in response to an incremental change in output voltage. The output resistance is then inversely proportional to the total collector current. To account for this effect, rather than including the output voltage dependence as an explicit resistive component in the circuit topology, the model extrapolates the slope on the I-V curves back to a common V_{CE} intercept known as the Early voltage. Then, and as shown in Figure 3.14, the Early voltage V_A models the slope of the output curves as if the output

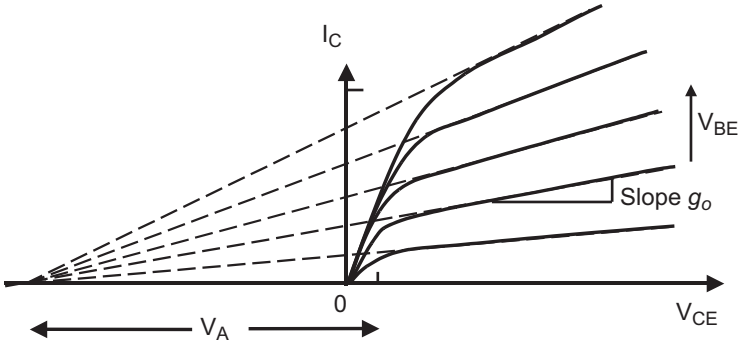


FIGURE 3.14
Output I-V curves for a bipolar transistor showing the definition of the Early voltage.

resistance becomes smaller with increasing base-emitter voltage [from (3.4) and because the slope of the I-V curves decrease with input voltage].

The resulting equations to represent the variation of all the model components are rather complex. The model topology is somewhat more straightforward, and the components for the NPN transistor model are shown in Figure 3.15.

Compared with the Ebers-Moll model, the complexity is immediately obvious. More than 30 parameters are now required to describe the equations representing the model, including factors to describe the behaviors noted above. The parasitic resistances, capacitance, and inductances are straightforward to specify (if not to measure); the nonlinear device capacitances are described by the capacitance at zero bias voltage as a reference point. Within the circuit topology itself, it is possible to generally compare the ideal diodes from the Ebers-Moll model with corresponding diodes, and their associated shunt capacitance, in the Gummel-Poon model; although two diodes are now needed for each junction to allow for the inflection point of the base current in Figure 3.12 with base emitter voltage, and its corresponding inflection point with base collector voltage. As for the Ebers-Moll model, the conduction current between the collector

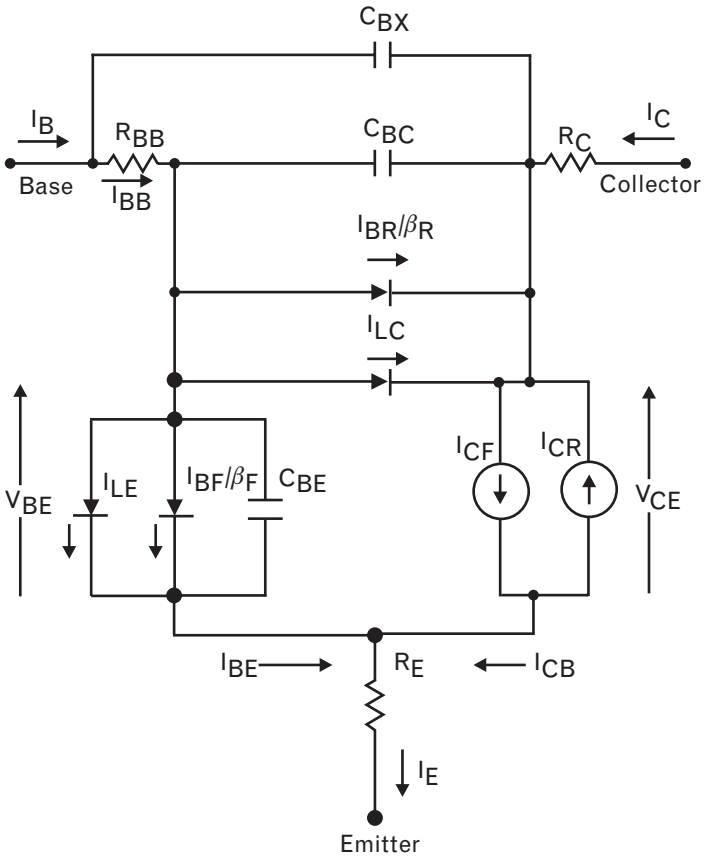


FIGURE 3.15
Equivalent circuit topology for the Gummel-Poon transistor.

and emitter is modeled by a forward and reverse current source that incorporates output resistance via the Early voltage. The emitter and collector resistances have been added, with the base spreading resistance R_{BB}. This resistor represents *current crowding* in the base as current injected normal to the plane of the base turns to spread within it. It is topologically included between two capacitances C_{BC} and C_{BX}, which represent the distributed nature of the space charge associated with the (generally) reverse-biased and somewhat long depletion region in the collector. In fact, the most common problem with the Gummel-Poon model is its poor representation of the transistor input (base) impedance at RF frequencies. This is associated with the inability of a single capacitance to represent electrons that are scattered through the base-collector junction. In fact, a series of R-C sections would be required to properly model this effect, rather than a single C-R-C section as with the C_{BX}, R_{BB}, and C_{BC} components in Figure 3.15. Numerous attempts to correct this problem have been attempted, perhaps the best known being to externally connect diodes between the base and collector as shown in Figure 3.16. Because the diodes are reverse biased, they contribute no conduction current but help to model the distributed nature of the charge and the resistance it encounters traversing the base-collector junction.

The PNP Gummel-Poon model is similar, although some care is needed to check the implementation within the CAD tool. Because the direction of the diodes is reversed to reflect the change from a p-n junction to an n-p junction, the direction of the current components and the control voltages across them is also changed. This can necessitate negative signs within some of the model equations, and the user should always check to see whether the CAD vendor has already accounted for these by changing the sign conventions within the model itself.

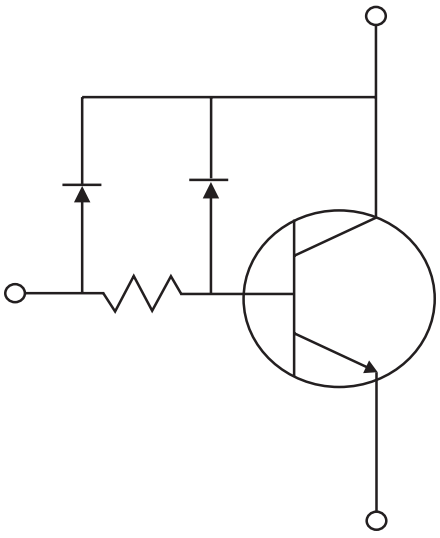


FIGURE 3.16
Modification to the Gummel-Poon model attempting to improve the accuracy of the input match.

Other problems with the Gummel-Poon model include errors in modeling the output conductance through the Early effect because of very small base widths; self-heating; the lack of modeling of the current that flows in the substrate; avalanche breakdown effects; and current crowding. A number of alternative models have been developed to overcome some or all of these shortcomings. They include the *vertical bipolar inter-company* (VBIC) model, developed by an industry consortium, and the Mextram model, developed by Philips. Both of these models are public domain models available on the Internet and have been implemented in some non-linear simulators. Unfortunately, parameter libraries for the transistors themselves are not yet widely developed.

3.2.2.5 Small-signal model derived from the Gummel Poon model

The small-signal T-model derived earlier from the Ebers-Moll model can now be recast in the form of the Gummel Poon model. Figure 3.17 shows the resulting hybrid-pi model, so named because of its topology.

This model includes from the T-model the base spreading resistance r_B and the capacitances associated with the base-emitter and base-collector diodes, referred to as C_{be} or C_π, and C_{bc} or C_μ. The latter sometimes can also include the parasitic package capacitance, although for the present, we have neglected other parasitic effects such as lead inductances and bonding pad capacitances.

The combination of the shunt input resistor and capacitor combination forms an RC time constant that sets the frequency of the dominant pole of the device (i.e., the pole that causes the current gain to roll off with frequency). Looking in at the device input (the base), the capacitance from both diodes appear in parallel when the output is shorted to calculate the frequency response of the short-circuited current gain; however, because the base-emitter diode is forward biased, C_π swamps the much smaller C_μ from the reverse-biased base-collector junction. If h_{fe} is the forward current gain of the device, and h_{fe0} its dc value—also equal to β_0—then it is relatively straightforward to derive

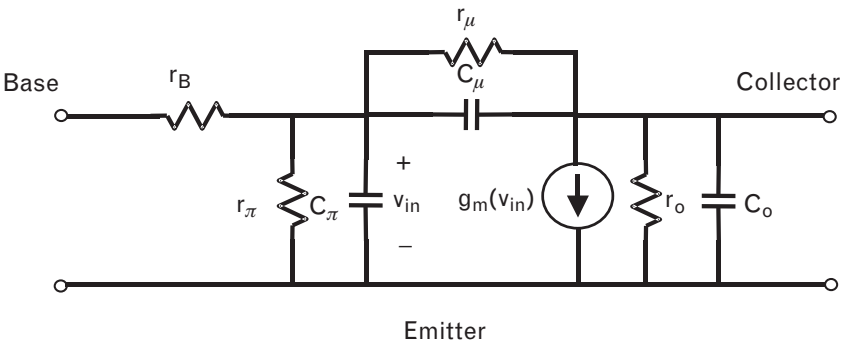


FIGURE 3.17
The small-signal hybrid-pi equivalent circuit derived from the Gummel-Poon model for the bipolar transistor.

$$h_{fe} = h_{fe0} / \left[1 + j\omega r_\pi \left(C_\pi + C_\mu\right)\right]$$
$$\left|h_{fe}\right| = h_{fe0} / \left[1 + \omega^2 r_\pi{}^2 \left(C_\pi + C_\mu\right)^2\right]^{\frac{1}{2}} \tag{3.22}$$

if we neglect the zero resulting from C_μ. We will refer to the frequency at which the *current gain* of the device is reduced by 3 dB from its dc value as f_{3-db}. It can be calculated from the above by setting the imaginary part of the denominator equal to one and solving for $\omega = 2\pi f$, that is,

$$\begin{aligned} f_{3-dB} &= \left[2\pi r_\pi \left(C_\pi + C_\mu\right)\right]^{-1} \\ &\approx \left[2\pi r_\pi C_\pi\right]^{-1} \end{aligned} \tag{3.23}$$

A similar expression for the frequency at which the *transconductance* of the device is reduced by 3 dB from its dc value can be obtained by replacing r_π in this expression with r_B. Even for high-frequency devices, f_{3-dB} is typically only in the tens of megahertz frequency range. Although this defines the frequency at which the gain begins to roll off from its dc value, the device can still have significant gain at higher frequencies. It is helpful to define the frequency f_T at which the current gain has fallen to unity by setting the numerator and denominator equal in (3.22):

$$f_T \approx h_{fe0} / \left[2\pi r_\pi \left(C_\pi + C_\mu\right)\right] \tag{3.24}$$

A physically based expression can be derived from the total emitter-collector transit time τ_{ec} discussed in Section 3.2.2.1, using

$$f_T = 1 / 2\pi\tau_{ec} \tag{3.25}$$

Equation (3.24) can be simplified by calculating the incremental resistance looking into the emitter of Figure 3.17, the so-called emitter resistor. This is calculated by applying an incremental voltage at the emitter and calculating its ratio to the resulting current

$$\begin{aligned} r_E &= \frac{\delta v_{EB}}{\delta i_E} \approx \frac{\delta v_{EB}}{\delta i_C} \\ &= \frac{1}{g_m} \approx \frac{kT}{qI_E} = \frac{26\,\text{mV}}{I_E(\text{mA})} \end{aligned} \tag{3.26}$$

where we have used (3.3) and (3.19). It may be further noted that

$$r_E = \frac{\delta v_{EB}}{h_{fe0}\delta i_B} \approx \frac{r_\pi}{h_{fe0}} \text{ so}$$
$$h_{fe0} = \beta_0 = \frac{r_\pi}{r_E} = g_m r_\pi \tag{3.27}$$

where we have used (3.21), noting that the total input resistance looking into the base is approximated by r_π since it usually dominates the base resistor r_B. Equation (3.24) then simplifies to

$$f_T = \frac{1}{2\pi r_E\left(C_\pi + C_\mu\right)}$$
$$= \frac{g_m}{2\pi\left(C_\pi + C_\mu\right)} \tag{3.28}$$

This transition frequency, or cutoff frequency, is a useful way to compare the upper frequency range of devices. It is typically in the gigahertz frequency ranges for most RF devices. Importantly, it depends on the collector current through g_m, so at low current values the device will have a lower f_T than at higher currents. It is also worth noting that if we define the *gain-bandwidth* product of a device *GB* as the product of its low frequency current gain and its 3-dB roll-off frequency, then

$$GB = h_{fe0} f_{3-dB} = \frac{h_{fe0}}{2\pi r_\pi\left(C_\pi + C_\mu\right)} = f_T \tag{3.29}$$

by substituting (3.23) and comparing it with (3.24). In other words, f_T is not only a measure of the frequency capability of a device, but also its low frequency current gain and its 3-dB roll-off frequency.

The measured behavior of f_T is illustrated in Figure 3.18. At low currents, f_T increases with bias current and g_m according to (3.28). There follows a range for which f_T is predominantly constant, and depends on the transit time a carrier takes to traverse the base and the collector space-charge region. The decline at high collector currents is due to an increase in the forward transit time of electrons across the base caused by high-level injection effects that enlarge the base region and cause lateral spreading.

However, operation of the transistor at higher frequencies is still possible because of the gain available due to the mismatch at the input and output. f_{MAX}, sometimes misleadingly called the *maximum oscillation frequency,* is measured when the maximum available unilateral power gain G_{MAX} becomes unity. Under these conditions, the device has been neutralized so there is no feedback. The current gain of a device is given by (3.22), while

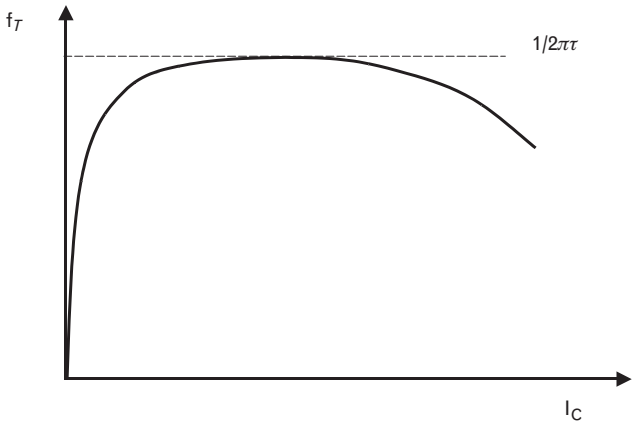


FIGURE 3.18
Behavior of the f_T of a bipolar transistor with collector current.

its voltage gain can be found by short-circuiting the input while loading the output by a current source. From these it can be shown that f_{MAX} is given by

$$f_{MAX} = \left(\frac{f_T}{8\pi r_B C_\mu} \right)^{\frac{1}{2}} \tag{3.30}$$

f_{MAX} is a better figure of merit for RF and microwave transistors because it depends not only on f_T but also on the parasitics of the transistor as well.

3.2.2.6 The common-base configuration

Most analyses of the transistor start with the common-emitter configuration. This is the most encountered transistor configuration for power transistors, because the fabrication process steps are optimized to achieve a low inductance ground there. The ground is also important as a thermal heat-sink for the device and is achieved without the need to electrically isolate the emitter from the heat sink.

The common-base configuration can be modeled using the same models as for a common emitter, but by redefining the input port to lie between the emitter and base and the output port between the collector and base. The resultant model is most easily seen by grounding the base in Figure 3.8. The input resistance of a common-base device is the much smaller r_E compared with r_π for a common emitter, so the location of the dominant pole at f_{3-dB} it forms with the emitter-base capacitance lies at a much higher frequency. Thus, the bandwidth of the common-base device is greater than for the common emitter, although the power gain is less because there is no current gain.

Any inductance in the common lead of the common-emitter transistor results in negative feedback, increase in input resistance, and reduction of

gain. However, the input and output voltages are in-phase in the common-base configuration, and any common lead inductance now results in positive feedback, a decrease in input resistance, and an increase in gain, often to the point of instability. This regenerative effect can be used on purpose to create wideband VCOs. In amplifiers however, the positive feedback due to regeneration tends to increase the nonlinearity compared with the common-emitter configuration.

The common-base configuration also tends to be more rugged to high VSWR loads than the common-emitter configuration. High VSWR loads create large peak voltages and currents because of reflection from the load. With common-base, the voltage at the collector can approach V_{CBO} rather than the lower V_{CEO} before stressing the device, since the base is well grounded at RF frequencies.

At higher frequencies, heterojunction bipolar transistors can also be used in common-base configuration. The common-base HBT is almost a unilateral device because there is very little feedback between collector and emitter and the input impedance is nearly resistive. These characteristics are perfect for the design of broadband power amplifiers with relatively flat gain. Of course, the requirement for a low inductance ground at the base is even more important at higher frequencies.

3.2.3 The heterojunction bipolar transistor

The HBT was a high-speed transistor first introduced commercially in the 1990s. AlGaAs-GaAs HBTs, sometimes called GaAs HBTs for simplicity, are fabricated using *p*-type GaAs in the base sandwiched between *n*-type GaAs in the collector and a layer of n-type AlGaAs in the emitter. The ohmic contacts to the emitter and collector are formed with heavily doped *n*-type GaAs. HBTs are also increasingly being constructed using InGaP (in the emitter layer) on GaAs and offer improved performance, including operation at lower battery voltages because of a lower turn-on voltage. In both cases, because GaAs is a poorer thermal conductor than silicon, thermal effects in these HBTs tend to dominate far more than for BJTs.

Using a compound semiconductor in the base creates a large potential barrier between the base and emitter and prevents holes being injected into the emitter from the base. This results in a higher current gain β since the base current is then theoretically reduced. More importantly, it means the base region can be more heavily doped than in a normal BJT and made thinner, decreasing the electron transit time and base capacitance, without increasing the base resistance. This results in better high-frequency performance.

The starting point for modeling the HBT is simply the bipolar junction transistor, since the behavior of the two devices is qualitatively similar. The HBT tends to be more linear in the active region than BJTs, with a value of

g_m that remains more linear with collector current. Comparing an otherwise identical BJT with an HBT in the Gummel plot of Figure 3.19, the higher current gain of the HBT is evident. Compared with the Gummel plot of Figure 3.12 characterized in the Gummel Poon model, *both* the base current and collector current reduce from ideal behavior in the high voltage region (region 3), and the reduction in β_F at high base voltages is no longer evident.

The dominant nonlinearities in any bipolar transistor or HBT are due to the forward-biased base-emitter junction and the nonlinearity of the base-collector capacitance (i.e., the reverse-biased junction). The effect of the latter is normally neutralized in fabrication by using either a thin collector layer or a low doping, so the junction remains fully depleted. Its capacitance is thus constant, at least until the device enters the saturation region where a large increase in base charge occurs and the capacitance rapidly increases. Regarding the former nonlinearity, in the HBT at least, the base-emitter junction conductance and capacitance show a fairly linear dependence on the base current in the active region, and only a weak dependence on the collector voltage. Using the transistor small-signal T-model at a number of bias points, it can be shown that the impedance parameter z_{12}, which relates the (total) input voltage to output current, is real and constant with frequency up to several gigahertz, principally because the base-emitter capacitance is very small. However, z_{12} is directly related to the emitter resistor and g_m, so is inversely proportional to the emitter current. Therefore, the HBT tends to be a very linear device with a high intercept point in a number of components. In relation to the 1-dB compression point, the third-order intermodulation distortion products can be lower than –20 dBc (20 dB below the fundamental) and the IP3 point well above 10 dB higher, even into the microwave frequencies.

In GaAs HBTs, the *low-current* recombination region (low base-emitter voltage) can be dominated by surface recombination of electrons

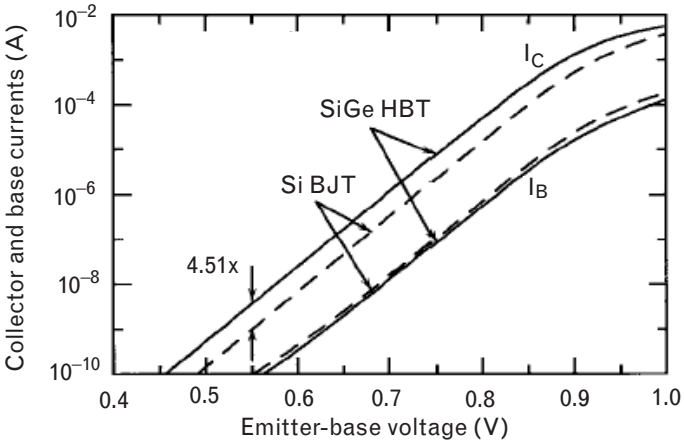


FIGURE 3.19
Gummel plot for the HBT (From: [5]. © 1998 IEEE. Used with permission).

injected from the emitter, and by leakage components. Because of surface states due to crystal lattice mismatch, the base current is nonideal at low bias voltages, and in some cases can even exceed the collector current. At *high currents*, several effects occur that are often difficult to separate [6]: thermal (self-heating) effects, the Kirk effect, and quasi-saturation effects. However, such high level injection effects are generally more benign in HBTs than in BJT's if self-heating is avoided (as in pulsed operation), because of the HBT's high base doping and lower current density. Self-heating is the rise in temperature due to power dissipation within the device. The Kirk effect occurs when the free charge in the collector exceeds the background doping level and the field at the base-collector junction vanishes. Driving the HBT beyond the onset of the Kirk effect will ultimately result in quasi-saturation, which occurs when a peak collector current density is reached that no longer depends on the base-emitter bias. This is the collector current density for which most of the base-collector bias voltage is dropped across the collector resistor R_C, and is temperature dependent. At this point, the base-collector junction is no longer reverse biased. Both effects ultimately cause the collector current to reach an ultimate maximum level.

Strictly speaking, the Gummel Poon model needs to be modified to account for these high-temperature and high-current effects. It also does not model the base-collector transit time for HBTs well. The increase in the small-signal transit-time as a result of these effects—typically by several picoseconds—can be incorporated into the Gummel Poon model, and will reduce the simulated gain by up to 1 or 2 dB, and thus the power-added efficiency. Without modeling for these effects, the HBT simulations give higher gain and will compress at a lower input power level. The even harmonics are also underestimated. These effects become worse when the operating frequency approaches the cutoff frequency of the HBT. Because of these deficiencies, the BJT Gummel Poon model (without modifications) is not always appropriate for HBT modeling under large-signal conditions near f_T. A number of models have been proposed in research papers, but at present there is no universally accepted HBT model.

3.2.3.1 The SiGe HBT

HBTs can also be fabricated using SiGe in silicon instead of AlGaAs in GaAs, and devices made in recent years have reduced the speed limitations normally posed by conventional silicon technology. The use of silicon provides better strength and thermal conductivity than GaAs, as well as better compatibility with CMOS technology. Processing can be done on 8-inch wafers so costs are also lower. The base-emitter turn-on voltage is also lower than for GaAs HBTs (0.8V versus 1.2V), allowing lower supply voltage operation.

In SiGe, between 10% to 20% of the silicon atoms in the crystalline silicon are replaced by germanium. By grading the doping of germanium through the base, a built-in electric field is induced across it that decreases the base transit time τ_b since carriers are now accelerated across the base. This increases the f_T of the SiGe HBT compared with a normal silicon BJT. The introduction of 90-nm line widths will further increase the operating speed of such devices.

In addition to mixed signal applications and integration with CMOS circuit functions, there has been much interest in using SiGe HBTs for the design of low phase noise oscillators, since the $1/f$ corner frequency can be in the range of 1-kHz. There is no noise performance penalty in doping the silicon with germanium compared with conventional silicon technology. Furthermore, although broadband noise figure results are not as good as with GaAs, they are certainly competitive (e.g., $F = 0.7$ dB at 2 GHz [5]). In fact, when coupled with their superior gain-to-dc power consumption ratio, a strong case can be made in favor of SiGe for most low-power system applications up to frequencies around 20 GHz [5].

The main drawback of SiGe circuits is the low breakdown voltage of silicon technology, restricting its ultimate output power. The SiGe HBT uses a moderately doped *p*-type base of SiGe sandwiched between a lightly doped *n*-type collector and an *n*-type emitter, both of silicon. For power devices, a thick collector with light doping is frequently used to increase the base-collector breakdown voltage and also to improve linearity, even though the resulting current density (and output power) will still be smaller than for the GaAs HBTs. Keeping the current density low helps avoid hot spots within the device that would otherwise render them useless for power performance. Higher power can also be achieved through stacking devices to increase the total current capability. In designing the SiGe HBT for higher power, a compromise has to be made between the requirements of breakdown voltage and f_T, since although the more lightly doped, or thicker, collector layer will improve the breakdown voltage, it increases the collector transit time. On one end of the scale, SiGe HBTs capable of 2-W output power with small-signal gains of 15 dB at 1 GHz have been fabricated. At the other end, devices with cutoff frequencies of over 100 GHz have been reported (with commensurately low breakdown voltage). The conventional Gummel-Poon model, ignoring thermal effects and taking into account only a change in Early voltage with bias and the Kirk effect, can suffice for good modeling of at least some SiGe HBTs [7].

In addition to SiGe and AlGaAs, HBTs are increasingly being fabricated in many other III-V materials, such as InP, InGaP, and InGaAsN. The InGaP HBT is particularly favored for handheld mobile operation because of its low knee voltage, sometimes as low as 0.5V, enabling it to operate with high efficiency from a 3-V supply rail. The current gain of InGaP HBTs is also far more constant with temperature than that of

AlGaAs HBTs, and they are easier to manufacture because their etching depth can be precisely controlled. They are also extremely linear, with IP3 points often more than 15 dB higher than their 1-dB compression points.

In summary, the HBT offers excellent gain, low noise, and linearity even at low-bias voltages. Unlike the MESFET or most HEMTs, the HBT also requires only a single dc source. These factors all contribute to the popularity of this technology.

3.2.4 The GaAs MESFET

The GaAs MESFET first became widely used in the 1980s and opened the way for the introduction of solid-state transistor techniques to microwave frequencies. Today, the MESFET and its heterojunction counterpart, the HEMT, are critical components in high-speed circuits of all types.

The cross-section of a simple MESFET is shown in Figure 3.20 [8]. The basic MESFET action arises because the metal gate fingers, lying on a slab of *n*-doped GaAs, form a metal-semiconductor, or Schottky junction. This can be modeled as a diode contact. The source and drain terminals are formed from highly doped areas at opposite ends of the gate. The *n*++ doping, used to indicate a heavy concentration of excess electrons at that point, forms an ohmic junction with the semiconductor. Unlike the Schottky junction, this junction is resistive in nature and supports free current flow in both directions across the contact.

The MESFET is called a *horizontal* device because electrons flow horizontally in the semiconductor channel between the source and drain, under the action of a large voltage differential between these two ohmic terminals. The current flow is supported entirely by free electrons within the *n*-doped GaAs channel. This makes the MESFET a majority carrier device, unlike the bipolar transistor where the current flow is due to minority carriers. The electron flow is modulated by the action of the gate, at which is applied a negative voltage that tends to deplete the region under the gate of electrons. As the gate voltage is made more negative, the area under the gate is ultimately depleted of all free electrons and the current is said to be pinched off. Only a leakage current component in the substrate

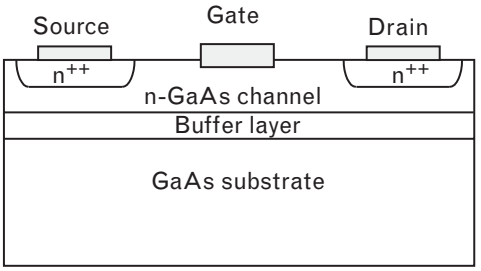


FIGURE 3.20
Simplified cross-section of a MESFET device.

remains at pinch-off. The corresponding gate voltage at this point is called the pinch-off voltage.

MESFETs gain their high speed compared to JFETs for a number of reasons. First, the devices are constructed with very small gate lengths (as low as 0.1 μm). This is the distance under the gate (from left to right) in Figure 3.20. The smaller the gate length, the faster the electrons can traverse the depleted region and the quicker the device responds to changes in gate voltage. Device capacitances are correspondingly smaller. Furthermore, the electrons in a semiconductor such as GaAs have very high mobility, corresponding to a high speed and good frequency response. Avoiding the use of minority carriers (holes in the case of a MESFET), which have much lower mobilities and slower speeds, and which increase the diffusion capacitance of the device, also enhances the device speed.

The gate width is the distance of the gate into the page in Figure 3.20. Increasing the gate width proportionally increases the transconductance of the FET and its current handling capability, or power. Unfortunately, because the gate is a thin strip of metal, it can also increase the gate resistance R_G, and thus the noise figure of the device. It also forms an R-C filter with the input Schottky capacitance of the gate that dominates the gain roll-off and can reduce high-frequency performance. Consequently, power MESFETs, with their high maximum channel current, will tend to have lower gain than small-signal devices. Although the gate resistance can be kept low by connecting a number of lower current cells in parallel, the shunt capacitance increases proportionally at both the input and output. A decrease in the equivalent series resistance at the input and output, and a lowering of gain, are inevitable consequences of this "gearing up" for current and power.

As the applied voltage at the drain increases, two effects occur. First, the velocity of the electrons increases until it eventually reaches a maximum, or saturated electron velocity. The current increases fairly linearly up to this point with increase in drain voltage. Second, the region between the drain and gate, itself a Schottky junction, becomes increasingly depleted as electrons are withdrawn from the affected area at the drain end of the gate, under the action of the strong longitudinal electric field. The depletion region forms an electric dipole between the drain and gate, one that is much stronger than the dipole between the source and gate because the voltage differential between the source and gate is weaker. This dipole ultimately limits the current that can flow between the drain and source since it narrows the drain end of the conducting channel due to depletion. As the drain voltage is increased further beyond this point, the incremental voltage is all dropped across the dipole, extending it further along the channel towards the gate, while the current changes only slightly. The device is then in so-called *current saturation* as the electrons are squeezed, at saturated drift velocity, into a narrow part of the channel between the depletion

region and the substrate. The knee voltage that separates the linear and saturated regions (to the left and right of the knee, respectively) corresponds to the drain source voltage at which the electrons reach saturated electron velocity and the channel under the gate is narrowed to the point where additional current cannot flow. This is a rather complicated function of the gate voltage, as we will see in the modeling of the MESFET drain current.

3.2.4.1 MESFET equivalent circuit model

Figure 3.21 shows the general topological model for the MESFET. Package effects, which introduce series inductance at the terminals and shunt capacitances between the terminals and ground, are omitted here for simplicity. At the core of the model are two diodes and a current source connecting the intrinsic terminals of the source, gate, and drain, in a similar way as for the Ebers-Moll model for the bipolar transistor. Both diodes have capacitances associated with them that represent the dipoles between the gate-source and gate-drain regions, which are typically reverse biased. The control voltage is V_{GS} and is measured across the gate capacitance. An intrinsic resistance R_I is included in series since the doping of the semiconductor presents a resistive component to the RF gate current flow. This, plus the metallization resistance R_G of the gate, and the parasitic resistance of the ohmic source terminal R_S, represent the resistance seen between the external gate and the source terminals. Like the gate resistance, the source resistance needs to be minimized to reduce the noise figure. Because it introduces negative feedback, it also reduces the gain, as does any series parasitic inductance in the source, which also increases the equivalent MESFET input resistance.

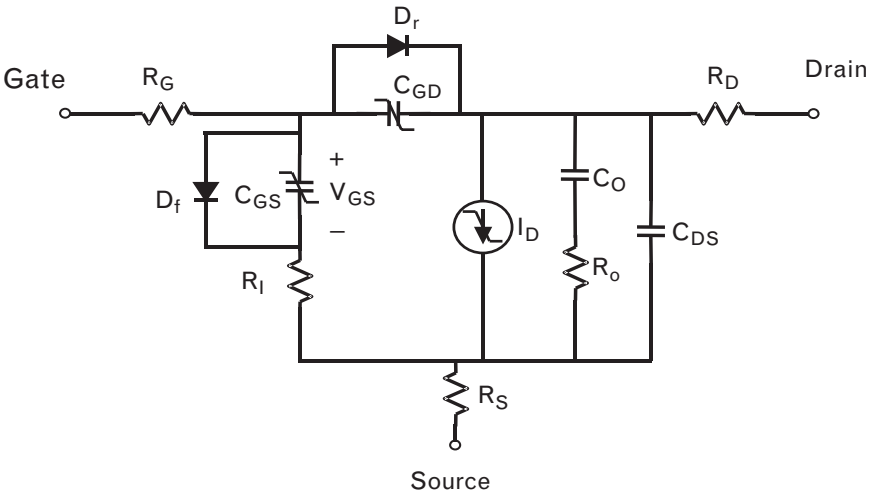


FIGURE 3.21
General MESFET model topology.

The current source represents the current flowing in the channel between the drain and source. This current source is controlled by the intrinsic gate voltage, and also by the drain-source voltage. In the saturated region, an ideal current source would have no dependence on the drain-source voltage if it had infinite output impedance. However, a MESFET has a finite incremental output resistance R_0, which is simply defined by (3.4) and should be built into the expression for drain current. This is not always the case and sometimes an additional output resistor between the drain and source needs to supplement the model. There is a further complication, however, and that is that the output resistance above a few megahertz in frequency can be much less than the dc value, which may be several thousand ohms. This is accounted for by the series combination of $R_0 - C_0$, which will decrease the output resistance to some value R_{DS} (implicit in the current source) in parallel with R_0 above the roll-off frequency determined by $R_0 - C_0$. The reduction in the small-signal high-frequency output resistance to around several hundred ohms is caused by impurities in the semiconductor lattice below the gate, which trap and release electrons into the current flow. The time constant of these traps typically corresponds to a few megahertz, so as the frequency is raised beyond this point the traps do not respond to the changing electric field and do not contribute to the steady-state current.

3.2.4.2 MESFET large-signal models

The Curtice model [9] is possibly the best-known MESFET model. It is available in most CAD tools, including SPICE. The model has a variety of implementations in the way it represents the drain-source current, but it relies on standard expressions to model the nonlinear behavior of C_{GS} and C_{DS}.

Most of the original work on MESFET models focused on obtaining expressions for the drain-source current, and accuracy of the model was staked on agreement between the measured and simulated current, as shown, for example, in Figure 3.22. The problem with this approach is that it neglects the other RF components of the model, and in particular, fails to model the decreased output resistance of the device at higher frequencies, known as dispersion of the FET output resistance. Of course, this effect occurs even for small-signal levels, and can be modeled by comparing the output resistance measured at dc with that derived from the *S*-parameters. A particular large-signal problem also arises because the I-V curves for many FETs actually show a negative output resistance at high currents and high drain voltages, indicated by the negative slope of the I-V curves in the saturated region for V_{GS} of 0V and –1V in Figure 3.22(a). Although it was postulated for a number of years that this effect could be due to the formation of oscillatory Gunn domains in the bulk GaAs channel, the effect is, in

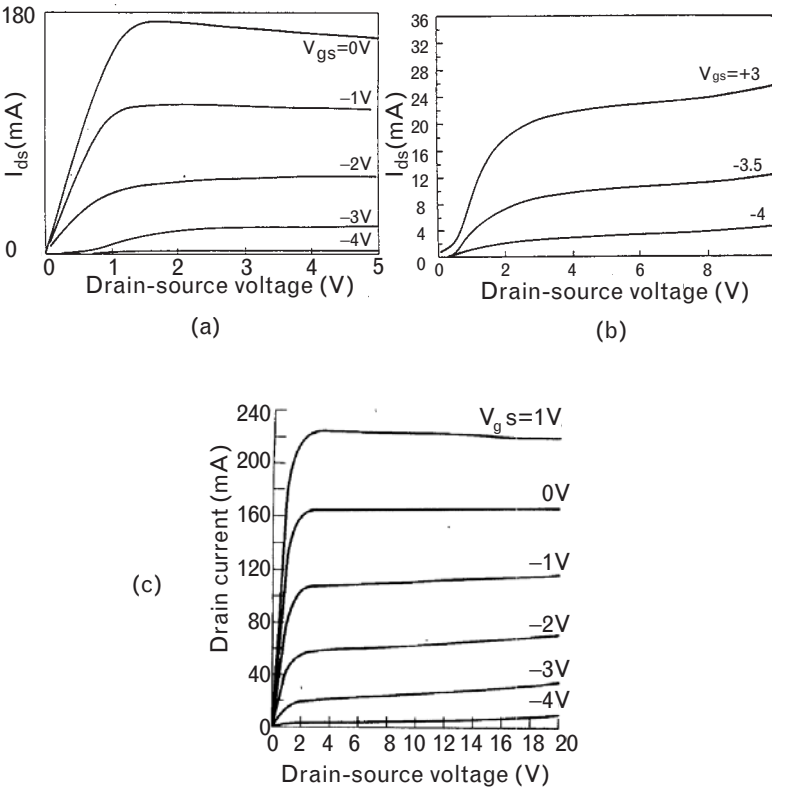


FIGURE 3.22
(a, b) Measured and (c) simulated I-V curves of an RCA MESFET using the Curtice cubic model. (From: [9]. © 1985 IEEE. Used with permission.)

fact, predominantly thermal, since at high drain currents the dc power dissipated in the device is quite high. This causes local heating, reduction in gain, and reduction in drain current compared to lower voltages where the temperature is cooler. A secondary effect that causes the apparent negative resistance is the low-frequency effect of traps. When pulsed measurements are made at frequencies of several megahertz, the negative resistance effect invariably disappears and steady-state I-V curves can be correctly modeled with positive output resistance. A third contributory effect is the bias dependence of the channel width.

The quadratic form of the Curtice model is given by

$$I_{DS} = \beta\left(V_{GS} - V_P\right)^2 \left(1 + \lambda V_{DS}\right) \tanh\left(\alpha V_{DS}\right) \tag{3.31}$$

The beauty of such a simple model is that its relationship to the measured quantity is straightforward and the components are easily derived. The first component is the classical square-law Shockley equation for a junction FET, relating the output current variation to the square of the difference between the gate voltage and the threshold (or pinch-off) voltage. Thus, β is the transconductance, since on differentiating (3.31) to obtain

incremental quantities β is directly proportional to the incremental drain-source current. The gate-source voltage, in fact, is usually modeled to incorporate a time-delay relative to the drain-source voltage, to account for the transit time of the electrons under the gate. The hyperbolic tangent models the general shape of the I-V curve: the linear region, the knee of the curve, and the flat current in the saturation region. Thus, α sets the knee voltage between the linear and saturated regions. The final variable λ sets the output resistance, since in the saturated region the hyperbolic tangent is approximately flat. Multiplication by a term proportional to the drain-source voltage adds the necessary slope to the I-V curve in this region.

Such a model is limited by its simplicity, since it forces g_m to be linear [simply differentiate (3.31) with respect to V_{GS}]. The Curtice cubic model removes this constraint at the expense of losing direct correspondence with measured results. The drain-source current is now modeled as a power series in a voltage that is a linear combination of V_{GS} and V_{DS} and has the form

$$I_{DS} = \begin{bmatrix} \left(A_0 + A_1 V_I + A_2 V_I^{\,2} + A_3 V_I^{\,3}\right) \\ +\left(V_{DS} - V_{DSDC}\right) / R_{DSO} \end{bmatrix} \tanh\left(\gamma V_{DS}\right) \tag{3.32}$$

where

$$V_I = \left(V_{GS} - I_{DS} R_S\right)\left\{1 + \beta\left(V_{DS0} - V_{DS}\right)\right\} \tag{3.33}$$

The term in A_3 now allows for nonconstant g_m. V_I is the intrinsic gate-source voltage across the internal MESFET junction at the reference voltage V_{DS0}, and β is a coefficient that changes the value of pinch-off voltage as V_{DS} departs from this reference voltage, to account for effects like substrate leakage. As for the quadratic model, γ sets the knee voltage between linear and saturated operation. R_{DS0} is the dc output resistance when V_I equals zero (i.e., when the drain current is at I_{DSS}).

There are a number of other formulations of drain current. For instance, Materka and Kacprzak use

$$\begin{aligned} I_{DS} &= I_{DSS}\left(1 - \frac{V_{GS}}{V_P}\right)^2 \tanh\left(\frac{\alpha V_{DS}}{V_G - V_P}\right) \\ V_P &= V_{P0} + \gamma V_{DS} \end{aligned} \tag{3.34}$$

while Statz, Smith, and Pucel use

$$I_{DS} = \frac{\beta\left(V_{GS} - V_P\right)^2}{1 + b\left(V_{GS} - V_P\right)}\left(1 + \lambda V_{DS}\right)\tanh\left(\alpha V_{DS}\right) \tag{3.35}$$

Further variants have emerged in recent years to improve the modeling accuracy of the drain current. The Chalmers model [10] offers improved accuracy over the Curtice model in the linear and saturation regions, while the Parker-Skellern model [11] and the Ooi model [12] are claimed to give accurate fitting in the knee and pinch-off regions as well. At some point, however, thermal effects and the dispersion of the output conductance limit the effectiveness of matching to measured dc characteristics, and the device will require more detailed measurements for accurate fitting.

As a user, the form of these equations is rarely important. The parameters are typically part of the device library supplied by either the manufacturer or the CAD vendor. The main limitations to keep in mind with all these models are that they can model the dc I-V curves rather well, but the capacitance variations and modeling of the output conductance at RF are rarely commented upon. Breakdown effects should also be modeled as part of the diode equations for D_r and D_f. To first order, breakdown is often modeled whenever the drain-gate voltage exceeds a fixed threshold, although in practice this also depends slightly on the gate-source voltage as well [13]. Breakdown can be an important effect in power amplifiers, and the designer needs to check if it is incorporated into the particular model being used.

3.2.4.3 MESFET small-signal model

The small-signal model for the MESFET in Figure 3.23 can be simply derived from Figure 3.21. The three principal nonlinear elements in that model are replaced by their linear equivalents, while the diodes, which are

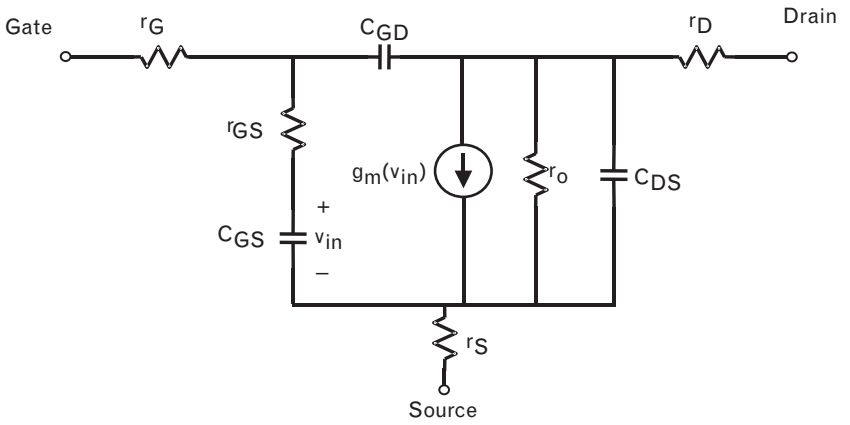


FIGURE 3.23
Small-signal equivalent circuit for the MESFET.

reverse biased, are usually omitted, although they can be modeled as very high resistances.

The output current source is replaced by a voltage-controlled current source, in exactly the same manner as for the bipolar transistor, where the controlling voltage is the voltage across the gate-source capacitance. The FET output resistance is modeled as a resistor r_o, usually different from that at dc. The resulting topology is remarkably similar to that of the bipolar transistor. However, because the diode between the two input terminals (gate and source) is now reverse biased, it is modeled as a series R-C circuit, with the gate, source, and intrinsic resistances forming the real part of the gate impedance that would typically be matched to 50Ω.

The frequency of unity current-gain f_T is defined analogous to that for the bipolar transistor,

$$f_T = \frac{g_m}{2\pi\left(C_{GS} + C_{GD}\right)} \triangleq \frac{1}{2\pi\tau} \tag{3.36}$$

where τ is the transit time as the electrons traverse the channel, and g_m is the dc value. The frequency at which the maximum available gain drops to unity is given by

$$f_{MAX} = \frac{f_T}{2}\sqrt{R_O/R_{IN}} \tag{3.37}$$

where R_{IN} is the series equivalent input resistance between the gate and the source, and R_O is the shunt equivalent resistance at the output between the drain and the source.

3.2.5 The high-electron mobility transistor

The classical HEMT is a MESFET-like device, and it usually refers to a device with an AlGaAs layer that forms the channel between the gate and the GaAs substrate. However, HEMTs are now increasingly based on a number of newer material structures, such as the (pseudomorphic) pHEMT (AlGaAs-InGaAs layers on GaAs), or the lattice-matched InP HEMT (AlInAs-GaInAs on an InP substrate). These devices all share the same horizontal structure of the GaAs MESFET but deploy additional layers to form a heterojunction in the channel under the gate. This results in a trapped layer of electrons (a two-dimensional "electron gas") having high saturated electron velocity under the compound semiconductor. This gives the HEMT its higher gain and extended frequency performance—applications well into the 200-GHz range have been reported. Their excellent low-noise properties result from their high transconductance and better electron mobility. Off-the-shelf HEMTs with noise

figures of 0.5 dB and associated gain of 12 dB at 15 GHz are available. Furthermore, by properly tailoring the HEMT layers, increased linearity can also result. Although the current density is lower and can limit the output power, multiple heterojunctions can be created to increase the power. As a result, the HEMT is gradually becoming the dominant FET technology for wireless applications.

As shown in Figure 3.24, most HEMTs show a characteristic peak in their transconductance versus gate voltage, with a convex shape. For negative gate-source voltages, g_m rises reasonably linearly; above that, g_m generally peaks between 0V and +0.5V on the gate, before starting to decrease due to forward conduction current. This gives rise to so-called enhancement mode pHEMTs, which are designed in such a way as to have a pinch-off voltage slightly above 0V, and to operate with gate voltages between that and the onset of forward conduction, around 0.6V or 0.7V. The pinch-off voltage is close to 0V and is then known as the threshold voltage. This removes the need for the negative bias rail typically required for MESFETs and normal depletion mode HEMTs and tremendously simplifies biasing since a simple resistive divider can be used from the drain supply rail. However, these devices have relatively low power density compared with HBTs, with which they compete, although they are inherently more efficient.

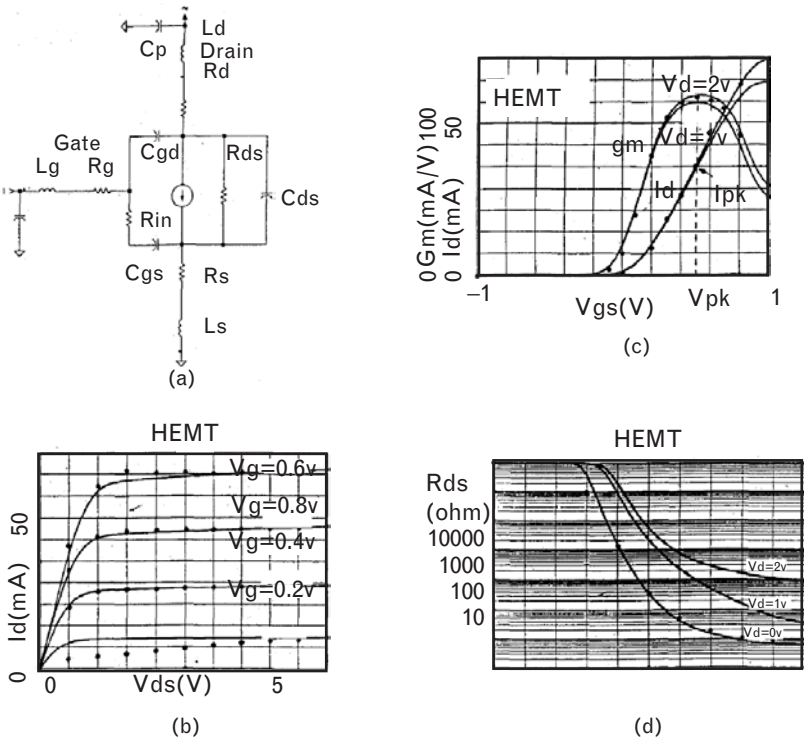


FIGURE 3.24
Measured characteristics of a pseudomorphic HEMT. (a) The equivalent circuit. (b) Drain current versus drain voltage. (c) Drain current and transconductance versus gate voltage. (d) Drain-source resistance versus gate voltage. (From: [10]. © 1992 IEEE. Used with permission.)

InP HEMTs in particular have g_m values of over 1,000 mS/mm of gate width. There is ongoing progress using them in the design of LNAs at ever-higher frequencies, with noise figures of a few decibels at more than 100 GHz. For example, [14] reports on an LNA operating at 183 GHz with a gain of 20 dB over a 30-GHz bandwidth, with a noise figure less than 5.5 dB.

The existing models already established for the MESFET [11, 12] are topologically identical to those for the HEMT, and can be used to model the HEMT's bell-shaped g_m curve. However, more complicated models specifically for the HEMT (e.g., as proposed by Angelov [10]), can be used to better model the characteristics of the capacitances C_{GS} and C_{DS}. Temperature-dependent elements account for self-heating, and the thermal time constant is modeled by an electro-thermal R-C equivalent circuit. The topology of the model, the transconductance, output I-V curves, and output resistance of a pseudomorphic HEMT are also shown in Figure 3.24.

MESFETs and HEMTs are also beginning to be made with new materials. Gallium nitride (GaN) and silicon carbide (SiC) are wide-bandgap materials that can handle high power densities, and they provide wide dynamic range. As a result, quite high operating drain voltages can be used in order to minimize the current necessary to achieve a given output power. Not only does this achieve higher output power, it can also prevent device impedances from becoming unreasonably low and as a result, high Q and narrowband.

Table 3.1 illustrates the basic material properties [15].

Table 3.1 Material Properties of SiC and GaN Semiconductors, in Comparison with Si, GaAs, and InP-Based Materials

Semiconductor/ Characteristic	SiC	GaN	Si	GaAs	InP
Bandgap (eV)	3.26	3.49	1.12	1.42	1.35
Breakdown field (MV/cm)	2.2 – 3.0	3.0	0.3	0.4	0.5
Electron mobility (cm^2/Vs)	700	1,000–2,000	1,500	8,500	5,400
Saturated electron velocity ($\star 10^7$cm/s)	2.0	1.3	1.0	1.3	1.0
Thermal conductivity (W/cm.K)	3.0 – 4.5	>1.5	1.5	0.5	0.7

Source: [15].

We can see that SiC, for instance, has a saturated electron velocity that is about two times higher than GaAs, as well as much better thermal conductivity. The combination of these properties may ultimately lead to high-frequency devices capable of generating and dissipating considerably higher power levels than GaAs MESFETs, although its limited carrier mobility will limit its range of frequencies. Power densities up to 10W/mm of gate width have been achieved with GaN at 10 GHz using an AlGaN/GaN HEMT, 10 times higher than for GaAs, and up to 7W/mm with SiC at 3.5 GHz. Because of their wide bandgap, such devices also have excellent radiation and heat resistance. For example, Cree Microwave has recently released a SiC device, the CRF-22010 FET, with a breakdown voltage of 120V and maximum operating temperature of 250°C. It can achieve 12-dB gain and 10-W output power at 2 GHz, when biased with 500-mA drain current at 48V drain bias voltage.

The electron mobility of GaN is better than SiC, but still lower than with GaAs, although the saturated electron velocities are comparable. Nevertheless, cutoff frequencies as high as 100 GHz have been achieved with HEMTs made from GaN. Noise figures below 1 dB at 10 GHz have also been reported. By adding a layer of AlGaN on top of the GaN, a two-dimensional electron gas is created in the GaN from the resulting heterostructure. This enables their exceptional high-frequency performance. GaN has amplified at ambient temperatures of 300°C. Their high breakdown voltage will also ultimately allow devices to be closely packed in integrated circuits.

Unfortunately, a number of production problems remain to be solved—the lattice structure of GaN is not well matched to most substrate materials, making them expensive to produce. Sapphire or SiC is most commonly used at present.

As with power HBTs and other power devices, self-heating effects cannot be ignored in modeling GaN HEMTs. The difference between pulsed and CW measurements often shows up in a "droop" or apparent negative resistance region in the output I-V curves, and lower gain at CW, due to the higher operating temperatures and consequent reduced mobility. Because of the large number of fitting parameters in the MESFET and HEMT models, they can also be applied to GaN HEMTs for nonlinear modeling. The small-signal model in Section 3.2.4.3 also gives a good fit at low power.

3.2.6 Silicon LDMOS and CMOS technologies

Because of the huge number of digital circuits fabricated using Si CMOS technology, it is natural that attention has turned to integrating RF circuits with the same technology, or the same process. Telecommunications technologies such as *synchronous digital hierarchy* (SDH) operate up to 10 Gbps

and wireless LANs at 100 Mbps, so RF techniques are essential even in the digital domain. The potential advantages of putting an entire system on a single chip are enormous in terms of cost, ease of fabrication, and perhaps even performance. Many low-frequency analog systems use digital processors on the same chip for auto-calibration and improvement of linearity. As the minimum feature size of CMOS has reduced to one-tenth of a micron, the speed capability of such systems has increased commensurately, and transistor cutoff frequencies above 30 GHz are becoming commonplace. However, integrating RF circuits into a single-chip system brings with it a host of problems as well, especially with noise, decoupling of the power supply, spurious products, implementing high-Q passive circuits, and reducing the device noise figure.

The CMOS FET is basically a square law device. Thus, its transconductance is given from (3.3) as

$$g_m = \frac{\partial I_{DS}}{\partial V_{GS}} = \frac{\partial \beta \left(V_{GS} - V_T\right)^2}{\partial V_{GS}} = 2\beta\left(V_{GS} - V_T\right) = 2\frac{I_{DS}}{\left(V_{GS} - V_T\right)} \tag{3.38}$$

But the denominator $V_{GS} - V_T$ determines the maximum input voltage swing of the device (i.e., essentially the input intercept point). Furthermore, the noise figure of the FET is inversely proportional to g_m. Thus, in an LNA, for example, once the noise figure and intercept point are specified, the drain current is fixed by (3.38). This is not necessarily a desirable constraint, so the receiver amplifier might be better implemented by a bipolar device. Bipolar transistors can also be more easily matched to each other than CMOS devices, and they exhibit lower noise and greater transconductance than a CMOS FET. More attention is thus required when using CMOS for precision circuits where tight control of amplitude and phase is required for differential signals, as in the quadrature upconverter. However, CMOS FETs produce excellent switches and have high input and output resistances, although they become quite inefficient at gigahertz frequencies. Thus, the combination of MOS and bipolar devices can result in an optimal system design.

BiCMOS technology using SiGe heterojunction bipolar transistors within the conventional silicon CMOS process will soon enable large-scale chips to integrate both RF and digital systems on the one chip. Such technology consumes less power and runs faster than all-CMOS designs. The only real drawback with silicon in higher frequency applications is the lack of material with which to fabricate low-loss transmission lines, and thus fabrication of high-Q inductors on chip still remains a problem to be

solved. Because the highly doped CMOS substrate creates a resistance to ground with the capacitance between the spiral turns of the inductor, Qs on even modified silicon substrates are usually less than 10. For comparison, monolithic spiral inductors on GaAs substrates have Qs in the range of 20 to 40. Nonetheless, the market is rapidly growing in favor of integrated SiGe front-ends rather than GaAs, and many of the ICs considered in Chapter 8 use SiGe within a CMOS-compatible process.

For very high-power RF designs, the *laterally diffused MOSFET* (LDMOS) is also a popular technology that can be fabricated in a CMOS process. It uses high resistivity silicon to achieve high breakdown voltages and the required output power. A *p*-doped layer of silicon is used underneath the *n*-type channel to isolate it from ground, and the source is grounded using *p*+ sinkers that run vertically from the source contact at the surface through to the bottom of the substrate. These low inductance "vias" improve the gain at higher frequencies, and the ability to directly ground the bottom of the substrate via a mounting flange means that the thermal conductivity can be kept high. They are also quite rugged devices and can withstand a high load mismatch at the input and output. Compared with MESFETs and HEMTs that require a negative bias at the gate, LDMOS requires only a single polarity supply since it is an enhancement mode FET. Its key drawback is its poorer linearity than its other FET counterparts. Compared with GaAs devices, their power added efficiency is also lower, and for this reason LDMOS FETs will probably be displaced in favor of GaAs pHEMT and HBT devices in handheld applications, where talk time is critical.

Silicon LDMOS and bipolar devices have dominated the market share of devices in the GSM cellular system, primarily because of their low cost rather than superior performance. Although primarily intended for power amplifiers up to about 2 GHz, they can also be used for power switching applications at lower frequencies. They are frequently used as power amplifiers in cellular base stations.

Standard digital models used for CMOS devices are not appropriate for RF design. In fact, the topology of the MESFET model provides a number of features for modeling the parasitic source, drain, and gate resistances and capacitances that need to be part of any RF MOSFET model. In particular, the gate resistance must be properly accounted for, particularly in modeling noise and transconductance. The effect of the substrate is also important; cross-talk can be accounted for by including a resistance network to model the bulk material under the active device.

In summary, silicon technology has many powerful incentives at RF, the greatest of which is its integration with baseband and digital systems. However, the physics embedded in the material properties of Table 3.1 also pose some fundamental constraints to silicon, which implies that there will always be multiple contenders for any application.

3.3 Problems

1. Show that the two models in Figures 3.7 and 3.9 are identical by comparing the terminal currents into the base, emitter, and current for each topology.
2. It would appear from (3.14) that I_{CT} approaches zero if the diodes are forward biased by an equal amount when the bipolar transistor goes into saturation. Can the emitter current component ever cancel out the collector current component? Under what condition will the total collector current entering the transistor equal the total emitter current entering the transistor? What is the base current then?
3. A bipolar transistor at room temperature requires 0.7V at its base to set the collector current equal to 1 mA. What is the value for I_S of the device, neglecting reverse leakage current? What would I_S be if the collector current were instead 100 mA for this base voltage? What would I_S be if the base voltage needed to be 0.75V to set the current to 1 mA?
4. (a) Putting FET cells in parallel is the usual way to obtain higher output current and achieve higher output power. If the gate is modeled as a series R-C circuit and the drain as a parallel R-C circuit, what is the equivalent input and output series resistance at the input and output for two cells in parallel, compared with one? What is the required transformation ratio of the input and output matching networks to 50Ω? What happens to the gain? At what frequency does G_{MAX} drop to 0 dB?

   (b) Repeat the above exercise, assuming now that the device is doubled in size by widening the gate. Assume that this doubles the input and output capacitance and shunt output conductance, but reduces the gate resistance to just three-quarters of its original value.

References

[1] Wei, C., et al., "Large-Signal Modeling of Self-Heating, Collector Transit-Time, and RF-Breakdown Effects in Power HBTs," *IEEE Trans. on Microwave Theory and Techniques*, Vol. 44, No. 12, December 1996, pp. 2641–2646.

[2] Gummel, H. K., and H. C. Poon, "An Integral Charge Control Model of Bipolar Transistors," *Bell System Technical Journal*, Vol. 49, May 1970, pp. 827–852.

[3] Antognetti, P., and G. Massbrio, *Semiconductor Device Modeling with SPICE*, New York: McGraw-Hill, 1988.

[4] Getreu, I., *Modeling the Bipolar Transistor*, Beaverton, OR: Tektronix Inc., 1976.

[5] Cressler, J. D., "SiGe HBT Technology: A New Contender for Si-Based RF and Microwave Circuit Applications," *IEEE Trans. on Microwave Theory and Techniques*, Vol. 46, No. 5, May 1998, pp. 572–589.

[6] Shirokov, M., et al., "Large-Signal Modeling and Characterization of High-Current Effects in InGaP/GaAs HBTs," *IEEE Trans. on Microwave Theory and Techniques*, Vol. MTT-50, No. 4, April 2002, pp. 1084–1094.

[7] Ma, Z., et al., "A High-Power and High-Gain X-Band Si/SiGe/Si Heterojunction Bipolar Transistor," *IEEE Trans. on Microwave Theory and Techniques*, Vol. MTT-50, No. 4, April 2002, pp. 1101–1108.

[8] Maas, S. A., *The RF and Microwave Circuit Design Cookbook*, Norwood, MA: Artech House, 1998.

[9] Curtice, W. R., and M. Ettenberg, "A Nonlinear GaAs FET Model for Use in the Design of Output Circuits for Power Amplifiers," *IEEE Trans. on Microwave Theory and Techniques*, Vol. MTT-33, December 1985, pp. 1383–1394.

[10] Angelov, I., H. Zirath, and N. Rorsman, "A New Empirical Nonlinear Model for HEMT and MESFET Devices," *IEEE Trans. on Microwave Theory and Techniques*, Vol. MTT-40, No. 12, December 1992, pp. 2258–2266.

[11] Parker, A. E., and D. J. Skellern, "A Realistic Large-Signal MESFET Model for SPICE," *IEEE Trans. on Microwave Theory and Techniques*, Vol. MTT-45, No. 9, September 1997, pp. 1563–1571.

[12] Ooi, B. L., J. Y. Ma, and M. Leong, "A Novel Drain Current I-V Model for MESFET," *IEEE Trans. on Microwave Theory and Techniques*, Vol. MTT-50, No. 4, April 2002, pp. 1188–1192.

[13] Fujii, K., et al., "Accurate Modeling for Drain Breakdown Current of GaAs MESFETs," *IEEE Trans. on Microwave Theory and Techniques*, Vol. 47, No. 4, April 1999, pp. 516–518.

[14] Rajit, R., et al., "183 GHz Low Noise Amplifier Module with 5.5 dB Noise Figure for the Conical Scanning Microwave Imager Sounder Program," *IEEE MTT-S Symposium Digest*, 2001.

[15] Eastman, L., and U. Mishra, "The Toughest Transistor Yet," *IEEE Spectrum Magazine*, May 2002, pp. 28–33.

CHAPTER 4

Nonlinear circuit simulation techniques

The history of circuit simulation for RF design has meandered back and forth between improvements to device models and improvements to the circuit simulators themselves. In the early 1970s, SPICE was pioneered and prompted a range of device modeling efforts, including the Gummel-Poon model for the bipolar transistor. Simple MESFET models were also introduced in the early 1980s, but the limitations of SPICE soon rendered further improvements to modeling accuracy wasted. It was not until the late 1980s and the commercialization of harmonic balance simulators for PCs that device modeling efforts once again accelerated, particularly for microwave monolithic integrated circuits and GaAs technology. The 1990s saw device manufacturers properly characterizing their devices and introducing device libraries, while a number of improved device models once again appeared. Now, in the early 2000s, PC simulation techniques have matured to the point where nonlinear characterization of diverse circuit phenomenon such as oscillation, phase noise, and high-order distortion products can be performed with relative ease. Perhaps the later years of this decade will again see the modelers playing catch-up, particularly for compound semiconductor devices.

This chapter will introduce the capabilities of nonlinear circuit simulators and help the user to understand the benefit that can be reaped from them. We will describe the principle differences between the simulators and when they might be used. We will not cover worked examples of nonlinear simulation until later chapters, when we set about designing specific components.

4.1 Classification of nonlinear circuit simulators

There are several different ways of analyzing a nonlinear circuit, and we will briefly examine each of these in turn.

4.1.1 Analytical methods

Analytical methods manipulate the algebra of the equations that describe a device. Such equations may be derived from the device physics or from curve fitting to measured data. For instance, the relationship between the 1-dB compression point and the third-order intercept point for a third-order nonlinearity is derived analytically for a given set of assumptions. The receiver spreadsheet analysis of Volume I, Chapter 3, deriving cascaded noise figure and intercept point, is an analytical technique.

Describing function techniques are another example, in which the input-output relationship is measured under known conditions of input voltage. The resulting ratio of output to input response, as a function of drive voltage, is known as the *describing function*. The describing function enables some analytical investigations, such as root loci and Nyquist plots for stability analysis of oscillators, or the design of an approximate match for power amplifiers.

Large-signal *S*-parameters are another example of describing functions that are sometimes useful. For simple nonlinear systems, the approach can be helpful to give a quick approximation of the system's performance, although its accuracy depends on the lowpass filtering characteristics of the embedding circuit because the harmonics from any nonlinear distortion are neglected.

4.1.2 Time-domain methods

With time-domain analysis, the state equations used to model the device behavior and its embedding circuit are expressed in the time domain. SPICE is the best-known time-domain analysis technique, and it is still widely used for both digital and analog analysis [1].

Consider the simple circuit of Figure 4.1 and assume the capacitor C is a nonlinear function of the applied voltage across it. We can analyze this circuit by first identifying the independent variables. For instance, we might choose these to be the voltage across the resistor $v_R(t)$ and the voltage across the capacitor $v_C(t)$. $v_R(t)$ and $v_C(t)$ are known as *state variables*. The dependent variables are the current in the resistor i_R and current in the capacitor i_C. They are related to the state variables through the relevant

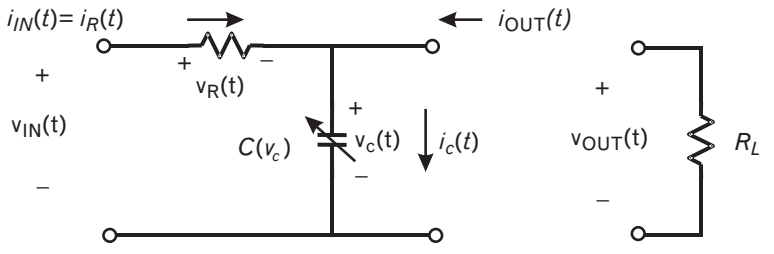


FIGURE 4.1
A simple R-C circuit in which the state variables are expressed in the time domain.

equations describing the behavior of their associated component. These equations are written in the time domain; thus,

$$\begin{aligned} i_R(t) &= \frac{v_R(t)}{R} \\ i_C(t) &= C(v_C)\frac{dv_C(t)}{dt} \end{aligned} \tag{4.1}$$

We can now apply Kirchoff's current and voltage laws to the circuit. Essentially, these are the topological constraints imposed by the way the circuit elements are interconnected. The first of these sums the total voltage around the input loop and imposes a constraint on the state variables allowing one of them to be removed from (4.1):

$$v_{IN}(t) = v_R(t) + v_C(t) \tag{4.2}$$

The second imposes a further constraint on the state variables by summing the total currents leaving the output node, assuming that the output load resistance is specified:

$$i_{OUT}(t) = i_C(t) - i_R(t) \tag{4.3}$$

The system of equations (4.1) to (4.3) is known as the state-space equations for the circuit since they are a system of equations that completely describes the behavior of the circuit in terms of the state variables and the circuit topology. If we can solve this system for the state variables for the imposed boundary conditions (e.g., the input voltage and output load), we can use Kirchoff's laws at the output node and output branch to derive the desired output quantities

$$\begin{aligned} v_{OUT}(t) &= v_C(t) \\ i_{OUT}(t) &= -\frac{v_C(t)}{R_L} \end{aligned} \tag{4.4}$$

Solving the state-space equations can be a nontrivial exercise, because (4.1) involves a derivative. Derivative equations are normally inverted to integral equations for numerical simplicity, although they require an estimate of the initial conditions from which to start integrating. In addition, we must evaluate the capacitance at each integration point by using the instantaneous value of the voltage across it at each time step during the integration summation.

In many RF designs, we are interested in the steady-state response of the circuit. Therefore, we must integrate over a sufficient number of RF

cycles to ensure that all transients have decayed and the circuit has reached steady state. This is one of the major drawbacks of the time-domain approach, because unless the initial conditions are well known, it might be necessary to integrate over many thousands of RF cycles for blocking capacitors or RF chokes to achieve their final steady-state voltage or current. In mixers or intermodulation analysis too, where the output of interest involves a component at possibly a much lower frequency than the RF frequency, numerous RF cycles have to be found before the low-frequency component can be extracted from the output (time-domain) waveforms.

What if the capacitor of Figure 4.1 were replaced with an open-circuited transmission line? In this case, we could split the voltage into an incident and a reflected voltage wave as shown in Figure 4.2.

If the time delay T for the wave to transit the line is known, then the boundary conditions could be represented as

$$\begin{aligned} v_1^+(t) &= v_2^-(t+T) \\ v_2^+(t) &= v_1^-(t+T) \end{aligned} \tag{4.5}$$

or similar equations if the line were not lossless. The terminal voltages are then simply given by

$$v_1(t) = v_1^+(t) + v_1^-(t) \tag{4.6}$$

$$v_2(t) = v_2^+(t) + v_2^-(t) \tag{4.7}$$

and can be used in the state-space equations as before. This is a very simplistic analysis because the time delay is rarely constant. In fact, microstrip transmission lines are generally dispersive (i.e., the effective dielectric constant is a function of frequency so the higher frequency components of a waveform travel more slowly than the lower frequency components). Although accurate spectral-domain models can be derived for microstrip to model this effect, such models are frequency-based by their very

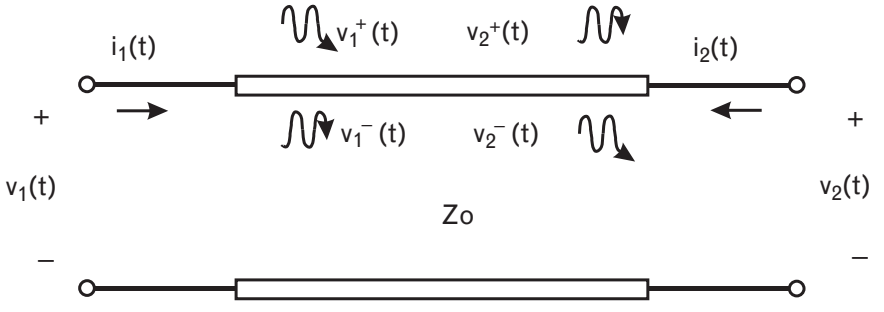


FIGURE 4.2
Modeling of a transmission line in the time domain.

derivation. Unfortunately, the concept of frequency has not at all entered our discussion of time-domain analysis—in fact, the concept of frequency cannot be applied in the case of a single pulse excitation, since it is not repetitive. This is both a strength and weakness of time-domain techniques. It is a strength because such a technique allows any type of waveform, periodic or not, to be analyzed, provided we can derive models for the circuit elements. Switches, oscillators at startup, chirp waveforms in radar pulses, and other sorts of waveforms that are nonrepetitive in nature can all be handled well by this technique. The weakness is that the concept of frequency is not intrinsic to time-domain simulation, and therefore, circuit elements that require a knowledge of frequency to represent them—as do most microstrip elements that are modeled by spectral-domain techniques—are not easily incorporated. Workarounds involving convolution of their impulse response to convert to and from the frequency domain are required.

As a final consideration, suppose that an additional capacitor C_{IN} is added in parallel with the input to the circuit of Figure 4.1. This now adds an additional equation to solve:

$$i_{C_{IN}}(t) = C_{IN}(v_{IN})\frac{dv_{IN}(t)}{dt} \tag{4.8}$$

where $i_{C_{IN}}(t)$ is the current through the added capacitor. This represents an additional computational overhead, since one more integration is now required, together with an estimate of the initial condition. The growth of the state-space equations with each additional node or branch added to the circuit makes time-domain analysis much less efficient than other means of analysis. This is the case even for linear components, because the overhead of integrating (4.8) remains even if the capacitance is a constant and not a function of voltage.

4.1.3 Hybrid time- and frequency-domain techniques—harmonic balance

The first harmonic balance simulators [2, 3] were introduced commercially in 1988, and since then they have become the workhorse of RF designers worldwide. Because they retained all the existing features of the traditional linear RF circuit simulators (including the ability to accurately model dispersive and distributed structures such as microstrip line and junctions) and frequency analysis with which RF designers were already comfortable, they were rapidly accepted as a design tool. Even today, with ever more powerful computing capability on the desktop, they have not been superseded by the more traditional time-domain analysis techniques such as SPICE. There are some fundamental reasons for this.

In harmonic balance, we make the a priori assumption that all waveforms within the circuit are periodic, or quasi-periodic (i.e., after a sufficient number of RF cycles all voltages and currents eventually return to the same values as at the beginning of the cycle). By doing this, in effect we have imposed on the input and output waveforms a representation of the form

$$v_J(t) = \mathrm{Re}\sum_{k=0}^{K} V_J(k\omega_0)e^{jk\omega_0 t} \tag{4.9}$$

where J is the node number of interest within the circuit, and there are up to K harmonics in the circuit about some fundamental frequency $\omega_0 = 2\pi f_0$, including dc when $k = 0$. This expression can be generalized in the case of mixers or intermodulation measurements to allow for more than one fundamental frequency, but it is simplest to consider the case of a single frequency excitation for now. Compared to time-domain analysis techniques, by enforcing (4.9) we have imposed the form of the solution on the state variables, and instead of searching for the waveform at all possible time points through numerical integration at each time point, we seek instead solutions only for the limited number of coefficients $V_J(k\omega_0)$ at each harmonic. These coefficients are just the phasor voltages at each frequency component present in the circuit. The solution set is then limited to finding the amplitude and phase of each phasor component, rather than an infinite number of time points. Of course, this has its limitations. We have in effect imposed a steady-state periodic solution on the circuit, thereby eliminating the analysis of arbitrary waveforms. Circuits such as switches, oscillators at startup, and amplifiers under transient conditions cannot be analyzed by the harmonic balance method. However, many RF components are, in fact, operated in steady-state; and even pulse-like driving waveforms can be represented as periodic waveforms if enough harmonics are included, so this limitation is not always severe. The advantages to be gained are immense:

- Dispersive, spectral-domain models for microstrip and other components can be incorporated because frequency is intrinsic to the harmonic balance analysis.
- Analysis times are faster, because we are solving only for a limited number of phasor variables to represent each state variable rather than a potentially unlimited number of time samples.
- Analysis time is independent of component values, because the voltage across a blocking capacitor or current in an RF choke, for example, is solved at steady state.

- Beat frequencies arising in mixers and intermodulation analysis add further state variables to be solved, but they do not require integration over possibly thousands of RF cycles for their extraction.
- Nonlinear models remain formulated in the time domain, consistent with their usual derivation from the semiconductor transport equations, which are functions of time.

There is an additional, more subtle advantage that arises from retaining the frequency domain within the analysis, and that is the same advantage enjoyed by linear RF simulators. Because the output phasor voltages and currents are related to each other by a linear impedance or admittance matrix containing all ports of interest, any internal nodes can be collapsed into a single matrix. This can result in considerable simplification of the analysis with implications for the speed of simulation and optimization. For instance, consider Figure 4.1. We can now rewrite (4.1) as

$$\begin{aligned} I_R\left(k\omega_0\right) &= \frac{V_R\left(k\omega_0\right)}{R} \\ I_C\left(k\omega_0\right) &= jk\omega_0 C\left(V_C\right) \end{aligned} \tag{4.10}$$

The differentiation has been replaced by multiplication by $jk\omega_0$, and the time-domain expressions for voltage and current have been replaced by their equivalent phasor values, shown capitalized for clarity. Equations (4.2) to (4.4) may also be rewritten as phasors to obtain a matrix representation:

$$\begin{pmatrix} V_{IN}\left(k\omega_0\right) \\ V_{OUT}\left(k\omega_0\right) \end{pmatrix} = \begin{pmatrix} R + \dfrac{1}{jk\omega_0 C} & \dfrac{1}{jk\omega_0 C} \\ \dfrac{1}{jk\omega_0 C} & \dfrac{1}{jk\omega_0 C} \end{pmatrix} \begin{pmatrix} I_{IN}\left(k\omega_0\right) \\ I_{OUT}\left(k\omega_0\right) \end{pmatrix} \tag{4.11}$$

which is an impedance matrix approach, or as

$$\begin{pmatrix} I_{IN}\left(k\omega_0\right) \\ I_{OUT}\left(k\omega_0\right) \end{pmatrix} = \begin{pmatrix} \dfrac{1}{R} & -\dfrac{1}{R} \\ -\dfrac{1}{R} & \dfrac{1}{R} + \dfrac{1}{jk\omega_0 C} \end{pmatrix} \begin{pmatrix} V_{IN}\left(k\omega_0\right) \\ V_{OUT}\left(k\omega_0\right) \end{pmatrix} \tag{4.12}$$

where the matrix is now an admittance matrix. Linear simulators already perform this task in order to calculate a circuit response, typically as a function of frequency. Now, as before, if we add another capacitance in shunt with the input, the rank of the impedance or admittance matrices in (4.11)

or (4.12) does not change. The element values will be affected, of course, but because the size of the matrix is unaffected, the time to perform the matrix multiplication in these equations and calculate each harmonic value of current or voltage is basically the same. The additional component we have added is simply collapsed into the same 2-by-2 matrix relating the input and output currents and voltages. Compared with the time-domain approach, where each additional component adds an additional state variable and requires its own integration, the speed advantage of the matrix approach is significant.

Of course, (4.11) and (4.12) are not quite as simple to solve as we have made them appear because the system of equations is nonlinear. In fact, the capacitance is a function of voltage, so the elements of the matrix in these equations actually depend on the state variables themselves. This is why we need to use the harmonic balance algorithm to find them. We will describe the algorithm shortly.

4.1.4 Frequency-domain techniques

The last approach is to represent the linear circuit in the frequency domain and to represent the nonlinear model in the frequency domain as well. This saves conversion between time and frequency, and it can be useful when large numbers of frequency components are present, as can occur in radio systems loaded with a large number of channels.

The Volterra series approach [4] is perhaps the best-known frequency-domain approach, although other variants exist such as generalized power-series [5]. Ironically for a frequency-domain approach, the nonlinear model starts out expressed as a power series in the state variables, and the device is assumed quasi-static so that the model represents the (time-domain) value of the output variables as functions of the instantaneous values of the state variables. The Volterra approach then extends the concept of an impulse response in a linear system, and its associated transfer function in the frequency domain, to a nonlinear system. If a system has an input signal $x(t)$ then the nth-order output response of the system $y_n(t)$ is formed from the convolution of the input with the nth-order nonlinear impulse response of the system $h_n(\tau_1,\ldots\tau_n)$:

$$y_n(t) = \int_{-\infty}^{\infty}\ldots\int_{-\infty}^{\infty} h_n(\tau_1,\ldots\tau_n)x(t-\tau_1)\ldots x(t-\tau_n)d\tau_1\ldots d\tau_n \qquad (4.13)$$

The impulse responses are known as the Volterra kernels of the system. The linear impulse response is simply $h_1(\tau_1)$, and the integration is then simply a single convolution integral of this impulse response with the input waveform. In the frequency domain, this corresponds to a multiplication. For higher-order terms the output is written similarly:

$$Y_n(\omega) = H_n(\omega_1 \ldots \omega_n) X_1(\omega_1) \ldots X_n(\omega_n) \tag{4.14}$$

$H_n(\omega_1, \ldots \omega_n)$ are known as the nth-order nonlinear transfer functions, analogous to the (first-order) linear transfer function $H(\omega)$. Having these as distinct functions allows the overall distortion to be deembedded into its individual nth-order components, where the total output is simply the sum of all the components in (4.13):

$$y(t) = \sum_{n=1}^{N} y_n(t) \tag{4.15}$$

To calculate the Volterra kernels is very messy. They are calculated in the frequency domain, in increasing order n. The circuit is represented by the usual series of interconnected elements, with unknown voltages at each node and currents and in each branch. The nonlinear elements are represented by power series of the respective voltage or current (whichever are taken as the state variables). The linear term of the response is then found in the usual way, by assuming the circuit is excited at a single frequency, calculating the output at the same frequency, and taking the ratio of input to output. For the nth-order kernel, the circuit is excited with state variables of the form

$$y(t) = \exp(j\omega_1 t) + \exp(j\omega_2 t) + \ldots + \exp(j\omega_n t) \tag{4.16}$$

that are substituted into the power series expressions for the nonlinearities. The resulting output current or voltage at the nonlinear component is then embedded into Kirchoff's equations representing the circuit topology. The terms of only the nth order are retained, so that when the ratio of input to output is taken, the time-domain terms in $\exp(j\omega_1 + j\omega_2 + \ldots + j\omega_n)t$ cancel at the input and output, and only the frequency-domain expression for $H_n(\omega_1, \omega_2, \ldots \omega_n)$ remains. For more details, the reader is referred to Maas [6] or Weiner [7].

These equations highlight the key reason why Volterra series is still used—because of its ability to handle mixing where multiple frequency components are present, and because intermodulation and mixing products of different orders, resulting from different components of the nonlinearity, can be analyzed and modeled separately. However, they also demonstrate that the process is very laborious, and practically only applicable to weakly nonlinear systems with analysis of at most third- or fourth-order terms.

4.2 The harmonic balance method

All the remaining nonlinear circuit design examples in this book use harmonic balance simulation. This is because amplifiers and mixers are inevitably measured, at least initially, using continuous-wave excitation, which is periodic by its very nature. Even oscillators are self-excited with a periodic waveform. Thus, we will now focus exclusively on the harmonic balance simulator, which is ideal for the design and optimization of RF components with this sort of drive. Only when startup and transient responses are required is it necessary to turn to a time-domain tool such as SPICE for closer inspection.

Figure 4.3 shows an arbitary single-transistor circuit deembedded into its linear and nonlinear parts. This enables us to define a clear split between those parts of the circuit that are analyzed in the frequency domain and those in the time domain. The linear circuit contains any matching networks, the bias network, device parasitics, sources, and so on. The nonlinear circuit contains only those elements within the device model whose value is a function of voltage or current. For instance, in the case of a MESFET, the nonlinear components might consist of the capacitance C_{GS} between the intrinsic gate and source, and C_{DS} between the intrinsic drain and source, in parallel with the drain current source. These elements are all modeled by equations in which the current through them is a function of the applied voltages at the intrinsic device terminals, perhaps with a transit time delay included, as well as possibly the derivatives of these voltages. The inclusion of a time delay means that the model need not necessarily be quasi-static (i.e., a function only of the instantaneous voltages and their derivatives). These voltages, $v_1(t)$ and $v_2(t)$ in the figure, are taken as the state variables of the system. In the more general case of multiple transistors, the number of nodes and branches joining the linear and nonlinear subnetworks will extend to some higher node number N rather than 2 as shown.

This is a convenient representation, because a model of the form $C = C(v)$, perhaps derived from the semiconductor physics of the device, or from empirical observation at a number of bias points, is assumed to imply $C[t] = C(v[t])$. The current is then $i(t) = C(v(t))dv(t)/dt$. More generally, the nonlinear currents in the branches joining the linear and nonlinear subnetworks can be modeled by equations of the form

$$i_j(t) = N\left(v_1(t), v_2(t) \ldots v_N(t)\right) \quad j = 1 \ldots N \tag{4.17}$$

where J is the branch number in consideration. We allow differentiation and integration in the equation to account for currents in nonlinear capacitors and inductors. N is any general nonlinear function of the state

variables—typical functions are power series, exponentials, and special functions such as Bessel functions.

In the case of Figure 4.3, the linear circuit can be modeled as a four-port network, in which two ports connect the linear and nonlinear subnetworks, and the other two are for applied bias and RF voltages. More generally, there will be $N + M$ ports, where M is the number of ports at which external sources are added. Such a linear network can be analyzed at each of the K harmonic components present in the circuit. The relationship between the linear applied voltages and the resulting currents at those ports is then an $N+M$-port admittance matrix at each frequency. Generally then, we can calculate an augmented admittance matrix at each frequency $k\omega_0$ where

$$\begin{pmatrix} \bar{I}_1(k\omega_0) \\ \vdots \\ \bar{I}_N(k\omega_0) \end{pmatrix} = \begin{pmatrix} Y_{11}(k\omega_0) & \cdots & Y_{1N+M}(k\omega_0) \\ \vdots & \ddots & \vdots \\ Y_{N1}(k\omega_0) & \cdots & Y_{NN+M}(k\omega_0) \end{pmatrix} \cdot \begin{pmatrix} V_1(k\omega_0) \\ \vdots \\ V_N(k\omega_0) \\ E_1(k\omega_0) \\ \vdots \\ E_M(k\omega_0) \end{pmatrix} \quad k = 0,1\ldots K \tag{4.18}$$

and the matrix is augmented from the normal square $N \times N$ admittance matrix to account for the additional M ports where external voltages are

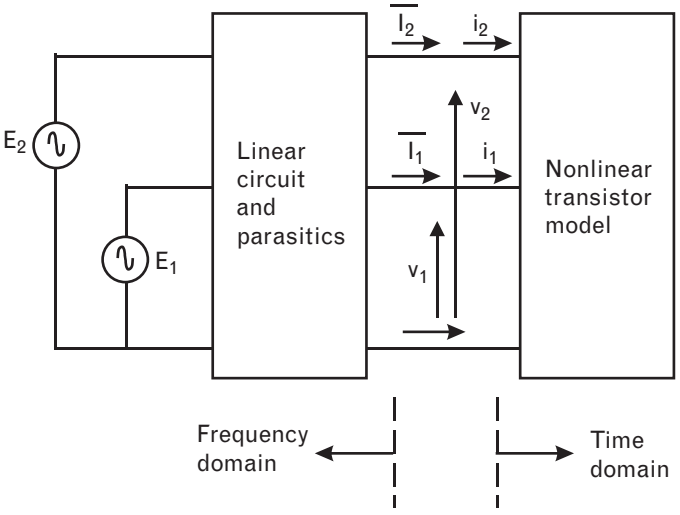


FIGURE 4.3
A simple transistor circuit deembedded into its linear and nonlinear parts.

applied. Most linear simulators already calculate the admittance matrix of a circuit in this form, although we should note in passing that the dc case usually requires special attention since the disappearance of inductors and capacitors at dc can result in singular matrices with many zero and infinite elements. The matrix in (4.18) is also usually required to be the *definite* admittance matrix, meaning that the ports connecting the linear and nonlinear subnetworks will not always be defined with one terminal as ground.

Returning now to Figure 4.3 for simplicity, we can describe the general process steps for the principles of harmonic balance. These are illustrated in Figure 4.4. For simplicity, let us assume that there are up to four harmonics of the fundamental frequency present in the circuit, generated by either an applied voltage or from harmonics created through distortion in the device.

1. Establish initial guesses for the frequency components of the state variables V_1 and V_2. If we let capitalized variables refer to phasor or frequency-domain quantities, then we establish initial guesses for

$$\begin{aligned} &V_1(0), V_1(\omega_0), V_1(2\omega_0), V_1(3\omega_0), V_1(4\omega_0) \\ &V_2(0), V_2(\omega_0), V_2(2\omega_0), V_2(3\omega_0), V_2(4\omega_0) \end{aligned} \tag{4.19}$$

These initial guesses should ideally correspond to the expected steady-state values of the state variables. In practice, the dc values probably correspond to the expected bias conditions, while the

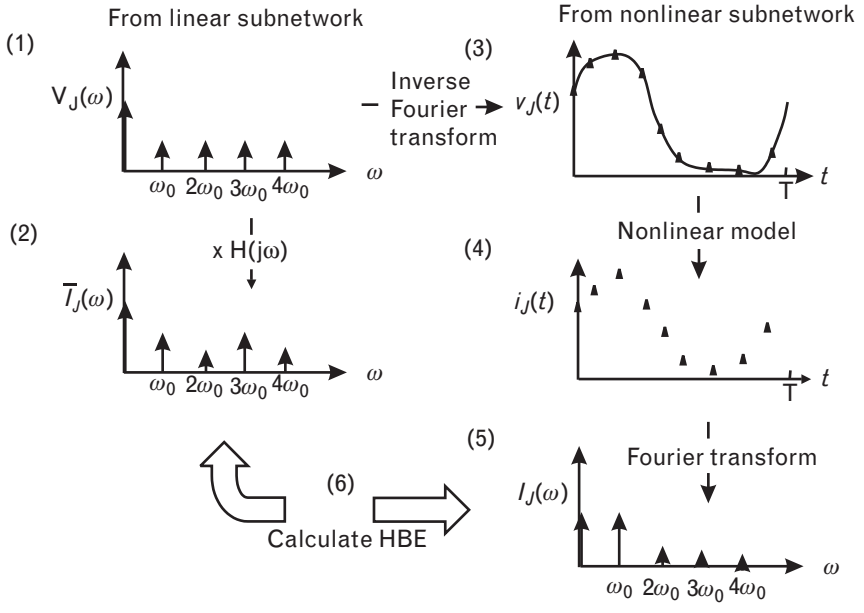


FIGURE 4.4
The harmonic balance process steps, for the circuit of Figure 4.3.

RF values can be initially set to zero since they are not usually known a priori.

2. Refer now to the linear subnetwork only. Use the values above for the state variables together with the known applied source voltages and the definite admittance matrix of (4.18) to calculate the corresponding values of phasor current that flow into the linear subnetwork, that is,

$$\begin{aligned}&\bar{I}_1(0), \bar{I}_1(\omega_0), \bar{I}_1(2\omega_0), \bar{I}_1(3\omega_0), \bar{I}_1(4\omega_0)\\ &\bar{I}_2(0), \bar{I}_2(\omega_0), \bar{I}_2(2\omega_0), \bar{I}_2(3\omega_0), \bar{I}_2(4\omega_0)\end{aligned} \tag{4.20}$$

The overbar is simply used to indicate the current flowing into the linear network.

3. Using an expression of the form

$$v_J(t) = \Re e \sum\nolimits_k V_J(k\omega_0) e^{jk\omega_0 t} \tag{4.21}$$

we can calculate the time-domain waveform corresponding to the two state variables v_1 and v_2. If required, the derivatives and integrals of the state variables can also be calculated directly from (4.21) by differentiation or integration. Since the waveforms are periodic, we need to evaluate (4.21) only at the Nyquist rate over one period (i.e., we must calculate the voltages at $2^{\star}K + 1$ time points within one period). In our example with four harmonics, we calculate nine time samples of v_1 and v_2 within one period T, that is,

$$\begin{aligned}&v_1(T/9), v_1(2T/9) \ldots v_1(T)\\ &v_2(T/9), v_2(2T/9) \ldots v_2(T)\end{aligned} \tag{4.22}$$

4. Now referring only to the nonlinear subnetwork, substitute the state variables from (4.22) and their derivatives or integrals into the nonlinear model (4.17), to yield values of nonlinear current $i_1(t)$ and $i_2(t)$ that flow at the same time instants, that is,

$$\begin{aligned}&i_1(T/9), i_1(2T/9) \ldots i_1(T)\\ &i_2(T/9), i_2(2T/9) \ldots i_2(T)\end{aligned} \tag{4.23}$$

5. Using a discrete Fourier transform, extract the frequency content of the time samples of current in (4.23). Since there are nine time samples, we can extract a dc component and four harmonics:

$$\begin{aligned} &I_1(0), I_1(\omega_0), I_1(2\omega_0), I_1(3\omega_0), I_1(4\omega_0) \\ &I_2(0), I_2(\omega_0), I_2(2\omega_0), I_2(3\omega_0), I_2(4\omega_0) \end{aligned} \tag{4.24}$$

6. Now putting the linear and subnetworks together, Kirchoff's current law is applied at each branch and requires that

$$I_J(\omega) = \bar{I}_J(\omega) \quad J = 1,2 \tag{4.25}$$

at all frequency components. We can calculate an error function comparing the components of current flowing into the two subnetworks as

$$\begin{aligned} HBE = &\sum_{k=0}^{K} \left| I_1(k\omega_0) - \bar{I}_1(k\omega_0) \right|^2 \\ &+ \sum_{k=0}^{K} \left| I_2(k\omega_0) - \bar{I}_2(k\omega_0) \right|^2 \end{aligned} \tag{4.26}$$

If we have reached a solution, then the current components at each branch and at every frequency component will be equal and opposite, and the *harmonic balance error* (*HBE*) will be zero. The method is called harmonic balance because the harmonics in the linear and nonlinear "sides" must balance each other out.

7. We return to step 1 and adjust the values of the state variables. The process steps above are successively continued until

$$HBE < \varepsilon \tag{4.27}$$

and the procedure is said to have converged. ε is typically of the order of 10^{-6} or smaller.

Once convergence is obtained, the solution to the state variables has been determined, and (4.18) can be used to find the branch currents. The linear subnetwork has therefore been solved. Quantities such as the distortion power, dc power, and gain can all be found through solution of the relevant currents and voltages within the linear subnetwork.

The harmonic balance procedure is thus an iterative procedure. Like all iterative procedures, there is no guarantee of convergence and even with today's "fail-proof" simulators certain circuits will have increasing values of *HBE* on successive iteration steps. The values of the state variables are usually adjusted using a quasi-Newton approach, in which slight adjustments are made in turn to each of the components of the state variables in (4.19), and the sensitivity of each of the resulting harmonic currents at each branch to that change can be calculated. In most cases, this involves the creation of what is known as the Jacobian, which is a sensitivity matrix, and

inverting that matrix. Since there are $(K + 1)$ unknown harmonic components at N ports, as the number of devices or harmonics increases, the time to find a solution increases quite rapidly.

Convergence can sometimes be achieved by adjusting some of the default parameters that control the harmonic balance engine. Increasing the number of harmonics to reduce the level of aliasing, decreasing the step size in the state variables between iterations, and sweeping the input power level from small-signal up to the desired large-signal level can all help approach the desired solution incrementally. The largest improvements in convergence, however, have come through mathematical tricks used within the simulators themselves. For instance, using the logarithm of the base voltage as a state variable for the bipolar transistor, rather than the base voltage itself, can help improve convergence since the base and collector current then vary linearly with that state variable rather than exponentially.

As can be seen from close inspection of the steps above, the algorithm can be applied generally to any circuit whose driving function, and thus response, is periodic, and whose nonlinearity may be modeled as a time-domain expression of the chosen state variables. Most RF amplifiers, mixers, attenuators, and filters fall into this class of circuit, as well as many systems. In the case of mixers, where both an RF and an LO signal provide two (usually) nonharmonically related input fundamental frequencies, a two-dimensional Fourier transform is required to support all possible linear combinations of these frequencies that are created within the circuit, including the IF. The same is required in the case of an amplifier to simulate its third-order intermodulation response. Most simulators also allow for a third fundamental frequency input, which is required for two-tone determination of the mixer RF response in order to simulate its third-order intermodulation performance.

4.3 Harmonic balance analysis of oscillators

We have discussed above the harmonic balance method for nonautonomous circuits, which are circuits with applied input signals. These input signals force the device into linear and nonlinear regimes that can be analyzed at known excitations and frequency. We will demonstrate the use of a harmonic balance analyzer to simulate mixers and amplifiers in the chapters that follow, since for these components the excitation and the output frequencies of interest are known.

In the case of an autonomous circuit, such as an oscillator, there is no applied RF signal and the frequency is initially indeterminate. Yet the solution is indeed periodic, thus should still be amenable to the harmonic balance approach. However, without a known frequency or excitation level, how can the state variables be driven to a steady-state value? There are a

number of different solutions to this problem. We will describe the theoretical basis for oscillator design using harmonic balance next, and leave the implementation until Chapter 6, where oscillator design is covered in greater detail.

4.3.1 Oscillator analysis using probes

A probe is a voltage source with series impedance, or a current source with shunt impedance, that is inserted into a circuit in order to drive the circuit in a forced regime. Ideally, it is attached to a node between the oscillating device and its resonant load. The probe is defined by its amplitude and fundamental frequency, and is assumed to have zero phase. Standard harmonic balance can then be used to analyze the circuit when the probe is inserted at a convenient point to drive the circuit, since it forces the excitation.

In order for the probe not to perturb the steady-state solution of the true circuit, the series impedance of the voltage probe is set to be infinite at all frequencies except the fundamental, where it is set to zero. Then, the voltage amplitude and frequency of the probe are adjusted so that at steady state the ratio of the probe current to its voltage equals zero at the fundamental. (The dual is true of the current probe.) Imposing this constraint on the converged solution implies that the probe can be removed from the circuit without affecting the result. The steady-state conditions will occur at that point where the circuit sustains its own excitation equal to the probe voltage and frequency.

Introducing the probe has now introduced two additional state variables (the fundamental frequency probe voltage and the fundamental frequency itself) into the harmonic balance system of equations. Both are initially unknown and need to be assigned initial values. However, the number of state-space equations that provide the boundary conditions has also increased by two as well, since the real and imaginary parts of the ratio of probe current to voltage must equal zero at the solution point. Thus, the system of equations remains square and can be solved using the same algorithm discussed earlier.

In actual implementation, the probe voltage and frequency are assigned initial values and an inner harmonic balance loop is first solved for the other state variables. A second, outer optimization loop is then used to adjust the probe voltage and frequency until the ratio of probe current to voltage equals zero. The initial value for the probe voltage is sometimes found by a separate search for that value of voltage that forces the loop gain to be one, which indicates the starting point is in the vicinity of oscillation. The nesting of two optimization loops is more demanding in terms of computer time but simpler to implement since it requires no modification to the harmonic balance engine for nonautonomous circuits. It also allows other variables to be associated with the probe and optimized as part of the

outer optimization loop, such as a tuning voltage or a component value that yields a desired oscillation frequency.

The probe type of analysis can also be extended to stability analysis of autonomous circuits [8].

4.3.2 Oscillator analysis using reflection coefficients of the device and resonant load

The analysis method described above is powerful and practical. However, it does not give any insight into how to design an oscillator circuit from scratch, or modify a circuit, for which the simulator finds no solution. If instead the oscillator design criterion is formulated as a stability constraint on the reflection coefficients of a device and its embedding circuit at the fundamental frequency, we can simulate the change in this design criterion as we vary the circuit.

This method of oscillator analysis [9, 10] can be applied to either linear analysis or nonlinear analysis of circuit stability. Since the required reflection coefficients can be obtained from *S*-parameters, the technique can also be used with a linear simulator to obtain good linear estimates of oscillation criteria. The same method can then applied with the additional use of a probe as described earlier to perform an equivalent analysis under large-signal or nonlinear conditions, using the full device model within a harmonic balance simulator.

The exact requirement for a feedback system to be unstable is that the poles of its closed loop gain function must lie in the right half plane. In Chapter 6, we will use the Nyquist stability criterion in such a system to assess whether it has such poles and hence determine its stability. However, it is not always easy to cast an oscillator in the form of a feedback system. At RF and microwave frequencies, it is sometimes easier to partition an oscillator circuit into an active device and a resonant load instead. We will see that the simplest way to analyze a circuit in this form is by considering the negative resistance of the active device as a function of the oscillating current or voltage.

Although the negative-resistance approach bears little if any resemblance to a closed-loop feedback system, one way to convert the negative resistance oscillator into a feedback system is by imposing a forcing voltage through a directional coupler, as shown in Figure 4.5.

For such a system, we can write

$$V^+ = \Gamma_D\left(V_i + V^-\right) = \Gamma_D V_i + \Gamma_D \Gamma_L V^+ \tag{4.28}$$

where Γ_D and Γ_L are the reflection coefficients of the device and resonant load, respectively. This may be rewritten

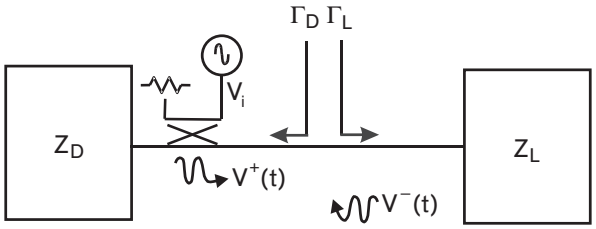


FIGURE 4.5 *A negative resistance oscillator split into its device and resonant load components, driven as a closed loop system by applying a forcing voltage through a directional coupler. (After:*

$$V^{+} = \frac{\Gamma_D(s)}{1-\Gamma_D(s)\Gamma_L(s)} V_i \tag{4.29}$$

which is just the equation for the closed-loop output of a feedback system in which $\Gamma_D(s)\Gamma_L(s)$ is the equivalent open-loop gain and $s = j\omega$ is the Laplace transform variable. The Nyquist plot for such a system is just the polar plot of this open-loop gain as a function of frequency. For this system, the Nyquist stability criterion states that the system will be unstable if the point +1 on the real axis is encircled at least once in a net clockwise direction by the polar plot of $\Gamma_D(j\omega)\Gamma_L(j\omega)$ as the radian frequency ω increases from negative infinity to positive infinity. Such a net encirclement would indicate that the closed-loop system has right-half plane poles.

This system is potentially unstable when the denominator of (4.29) is zero, or when $\Gamma_D(s)\Gamma_L(s) = 1$ in both amplitude and phase. We will see in Chapter 6 that this is equivalent to the well-known condition for steady-state oscillation requiring that the device impedance is equal and opposite to the resonator load impedance. However, this condition states nothing about oscillator *startup*, which requires right-half plane poles to exist close to the oscillation frequency.

To assess oscillator startup, we can use the same Nyquist technique as long as Γ_D itself has no right-half plane poles (i.e., if the device is not oscillating during measurement or simulation of its reflection coefficient). If this is the case, we simply make a polar plot of $\Gamma_D(s)\Gamma_L(s)$ using either a linear or harmonic balance simulator, with $s = j\omega$. To simulate the product of Γ_D and Γ_L, we can either build a circulator model as shown in Figure 4.6, or use the reflection gain probe built into some simulators, such as the GPROBE element in AWR's Microwave Office. This was first introduced in Chapter 1 when we analyzed the stability of amplifiers. The GPROBE element plots the negative of $\Gamma_D(j\omega)\Gamma_L(j\omega)$, so the Nyquist criterion must then be expressed in terms of encirclement of the point –1, rather than +1. This is, however, identical to what we are doing here, since the sign in the denominator of (4.29) is simply changed by plotting the negative of

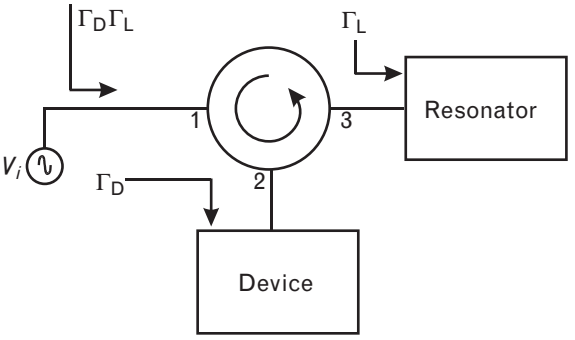


FIGURE 4.6
Evaluation of the product of the device and load reflection coefficients using an ideal circulator. (After: [9].)

$\Gamma_D(j\omega)\Gamma_L(j\omega)$, and the denominator now goes to zero when this (inverted) open-loop gain expression equals –1 rather than +1.

If we use the circulator implementation, the reflection coefficient at port 1 for an applied input voltage V_i is just $\Gamma_D(j\omega)\Gamma_L(j\omega)$, which can be plotted under either linear or nonlinear conditions for frequencies from 0 to ∞. This allows us to plot one-half of the Nyquist plot; the other half for negative frequencies is the mirror image of the first curve reflected about the x-axis, since $\Gamma_D(j\omega)\Gamma_L(j\omega) = \Gamma_D^*(-j\omega)\Gamma_L^*(-j\omega)$.

To see how this technique could be applied, consider the two circuits given by Jackson [9] and reproduced in Figure 4.7. In *both* cases, the device reflection coefficient at resonance is

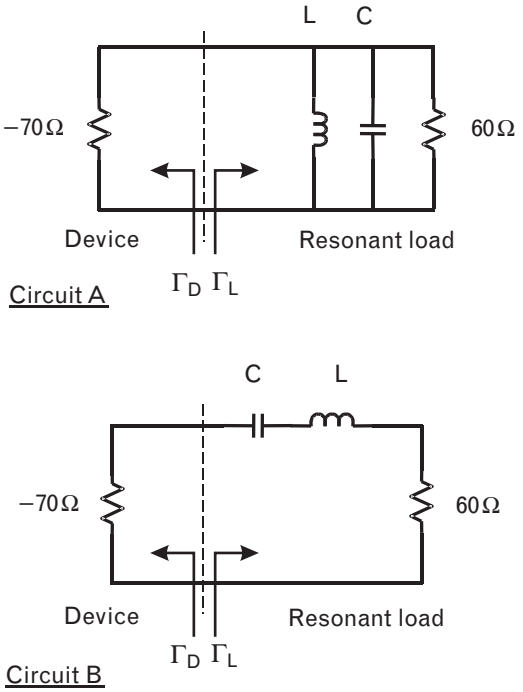


FIGURE 4.7
Two simple circuits with negative resistance: (a) a shunt circuit, and (b) a series circuit. (After: [9].)

$$\Gamma_D = \frac{Z_L - Z_0}{Z_L + Z_0} = \frac{-70 - 50}{-70 + 50} = 6 \tag{4.30}$$

and the load reflection coefficient at resonance is

$$\Gamma_L(j\omega_0) = \frac{60 - 50}{60 + 50} = 0.091 \tag{4.31}$$

Although both circuits have negative resistance or negative conductance, circuit A has left-half plane poles, so it is, in fact, stable. We will see in Chapter 6 that circuit A fails to satisfy the conditions for steady-state oscillation because the device conductance is not sufficiently negative when compared with the load conductance. On the other hand, circuit B has right-half plane poles and is unstable, and it will, in fact, sustain steady-state oscillation because the magnitude of the device resistance is greater than the load resistance, and is negative. Furthermore, from the two equations above, *neither* circuit satisfies the oft-quoted condition for oscillation buildup:

$$\left|\Gamma_D(j\omega_0)\right\|\Gamma_L(j\omega_0)\right| > 1 \tag{4.32}$$

This test fails to predict oscillation for circuit B. This latter condition is generally a poor criterion to use for assessing instability, since it implies a circuit loaded by 50Ω (with $\Gamma_L = 0$) can never be unstable.

In fact, only the Nyquist stability criterion can accurately predict the startup of oscillations. Of the two circuits in Figure 4.7, the Nyquist stability criterion correctly assesses instability for circuit B from $\Gamma_D(j\omega)\Gamma_L(j\omega)$. The polar plots for the two circuits are shown in Figure 4.8.

The Nyquist plot for circuit A does not enclose +1 at all, and is therefore stable. For circuit B, even though we see that $\Gamma_D(j\omega)\Gamma_L(j\omega) < 1$ at the zero-phase crossing, the plot still encircles +1 clockwise (twice) as the frequency goes from $-\infty$ to $+\infty$. This circuit is therefore unstable because it has two right-half plane poles. This example clearly illustrates that (4.32) alone is neither a *necessary* nor *sufficient* condition for oscillation.

Jackson [9] expresses the condition for clockwise encirclement of the point (1,0) mathematically, relating it to the slope of the $\Gamma_D(j\omega)\Gamma_L(j\omega)$ curve at the zero-phase frequency ω_0, corresponding to the presumed oscillation frequency. He shows that (4.32) is, in fact, a *sufficient* condition to predict instability and build up of oscillations only if the frequency slope of the reactive part of $\Gamma_D(j\omega)\Gamma_L(j\omega)$ at resonance is negative. The condition can be expressed graphically by referring to Figure 4.8 and considering

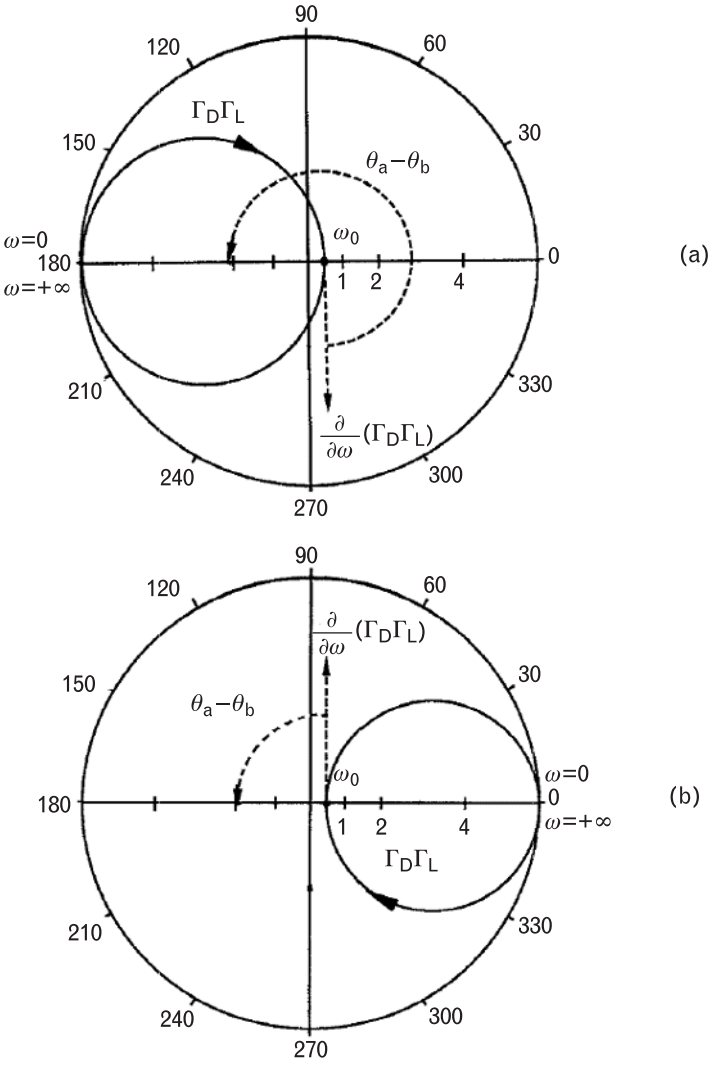


FIGURE 4.8
(a) Nyquist plot for circuit A in Figure 4.7, not circling +1 thus indicating a stable circuit. (b) Nyquist plot for circuit B circling +1, thereby indicating instability. (From: [9]. © 1992 IEEE. Used with permission.)

the angle between the tangent to the $\Gamma_D(j\omega)\Gamma_L(j\omega)$ curve at ω_0 (i.e., the reactance slope), and the vector from the point (1,0) to this curve at ω_0. If this angle (measured vectorially; i.e., counterclockwise) lies between 0° and 180°, the circuit has a right-half plane pole near ω_0 and oscillations will build up at that frequency. This is indicated by the angle $\theta_a - \theta_b$ in Figure 4.8.

In fact, (4.32) alone is irrelevant for instability; what is important is whether $\Gamma_D(j\omega)\Gamma_L(j\omega)$ encircles the point (1,0) in a net clockwise direction on the Nyquist plot. The magnitude itself can be either greater or less than one at resonance. For a single pair of complex poles this is equivalent to stating net encirclement of (1,0) by $\Gamma_D(j\omega)\Gamma_L(j\omega)$ and an angle as defined above that lies between 0° and 180°. Thus, for circuit B, the point (1,0) is encircled clockwise (and the angle is 90°), so right-half plane poles

exist and will guarantee start up of oscillations, even though $\Gamma_D(j\omega)\Gamma_L(j\omega) < 1$ *at resonance*. In this case, the slope of the reactive part of $\Gamma_D(j\omega)\Gamma_L(j\omega)$ with frequency at resonance is positive. For circuit A, there is no encirclement, so that circuit will be correctly judged to be stable.

Time-domain methods such as SPICE have traditionally been used to determine whether an oscillator will start up or not, since the initial conditions in SPICE can be set to zero and the algorithm supports full transient analysis, as illustrated, for instance, in [1]. Harmonic balance methods, on the other hand, already assume a periodic excitation exists within the circuit. Nevertheless, although the startup waveform cannot be derived, a harmonic balance simulator can still be used to predict whether the conditions for startup are satisfied, as shown earlier.

4.3.3 Oscillator analysis using a directional coupler to measure open-loop gain

An approach quite similar is to insert a four-port directional coupler into the feedback path of an oscillator to determine the open-loop gain directly. Such an alternative would be preferable to the analysis of reflection coefficients we used above if the feedback loop were clearly identifiable, or if driving a 50-Ω load with reflection coefficient close to zero.

Microwave Office, for instance, uses the OSCTEST directional coupler element, which breaks the loop only at the fundamental frequency and allows other harmonics to pass through unimpeded. This element is illustrated in Figure 4.9. It is a special directional coupler that allows us to inject a fundamental frequency input voltage into the system (port 3 in the figure), and to sample the resultant system output (port 4). The coupler prevents propagation of the generated fundamental (oscillatory) component back into the system itself (i.e., it keeps the system open-loop). Then, the loop gain can be tuned to achieve a phase of zero and gain magnitude greater than one at small signal levels, so that the conditions for steady-state oscillation are met.

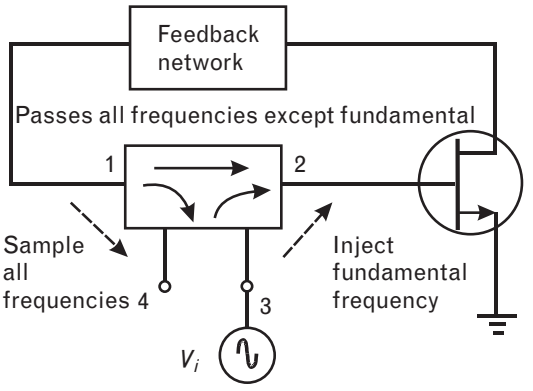


FIGURE 4.9
The directional coupler element used to measure the open-loop gain of an oscillator.

This element can also be used in an identical manner to that described in the section above. We can construct a Nyquist plot of the loop gain and check for clockwise encirclement of (1,0) to ensure the existence of right-half plane poles and oscillation startup. A *necessary* (but not sufficient) condition for oscillation startup is that the loop gain be greater than one at small signal, because as oscillations build up, the device will always compress. This condition needs to apply at some frequency close to the expected oscillation frequency. As the device begins to saturate, the loop gain will decrease to one at the steady-state oscillation frequency (the Barkhausen criterion). The total phase of the loop gain also needs to be a multiple of 360° at the oscillation frequency. This test on loop gain is more useful than the equivalent test on reflection coefficient described by (4.32) since it a necessary condition. The *sufficiency* is ensured by checking that the Nyquist condition [clockwise encirclement of (1,0) with frequency] is also satisfied. Because the circuit is driven by a signal at one port of the coupler, the conditions for oscillation may be checked at both small- and large-signal levels.

We will have more to say about oscillators and their startup and steady-state oscillation requirements in Chapter 6, with some examples. Here, we have introduced how the harmonic balance method can be used for the design and analysis of nonlinear circuits, even for those that are autonomous. First, however, in the next chapter we will use the harmonic balance method to design and analyze the most common class of *non*autonomous circuits, the power amplifier.

References

[1] Kundert, K. S., *The Designer's Guide to SPICE and Spectre*, Boston, MA: Kluwer Academic Press, 1995, Chapter 4.

[2] Gilmore, R. J., "Nonlinear Circuit Design Using the Modified Harmonic Balance Algorithm," *IEEE Trans. on Microwave Theory and Techniques*, Vol. MTT-34, No. 12, December 1986, pp. 1294–1307.

[3] Rizzoli, V., A. Lipparini, and E. Marazzi, "A General-Purpose Program for Nonlinear Circuit Design," *IEEE Trans. on Microwave Theory and Techniques*, Vol. MTT-31, September 1983, pp. 762–770.

[4] Volterra, V., *Theory of Functionals and of Integral and Integro-Differential Equations,* New York: Dover, 1959.

[5] Gilmore, R. J., and M. B. Steer, "Nonlinear Circuit Analysis Using the Method of Harmonic Balance—A Review of the Art," *International Journal of MW and MMW Computer-Aided Engineering*, Vol. 1, Nos. 1 and 2, Wiley, Part I – January 1991, Part II – April 1991.

[6] Maas, S., *Nonlinear Microwave Circuits*, New York: IEEE Press, 1997.

[7] Weiner, D. D., and J. F. Spina, *Sinusoidal Analysis and Modeling of Weakly Nonlinear Circuits*, New York: Van Nostrand, 1980.

[8] Suárez, A., J. Morales, and R. Quéré, "Synchronization Analysis of Autonomous Microwave Circuits Using New Global-Stability Analysis Tools," *IEEE Trans. on Microwave Theory and Techniques*, Vol. MTT-46, No. 5, May 1998, pp. 494–496.

[9] Jackson, R. W., "Criteria for the Onset of Oscillation in Microwave Circuits," *IEEE Trans. on Microwave Theory and Techniques*, Vol. MTT-40, No. 3, March 1992, pp. 566–569.

[10] Jackson, R. W., "Comments on 'Criteria for the Onset of Oscillation in Microwave Circuits'," *IEEE Trans. on Microwave Theory and Techniques*, Vol. MTT-40, No. 9, September 1992, pp. 1850–1851.

CHAPTER 5

High-power RF transistor amplifier design

It is quite common for new designers to wonder exactly at what point a small-signal device ceases to be classified as "small signal" and enters the realm of "large signal"—or, in fact, whether the distinction needs to be made at all.

Although there is no clear delineation, the distinction is quite important because prior to 1985, large-signal analysis was rarely applied to circuits because of its difficulty and the lack of simulation tools. The wealth of analysis we have covered so far, which centers around *S*-parameter analysis and equivalent linear characterization of devices, is so tractable and amenable to application that even today we attempt to apply this body of knowledge to components that are clearly large signal in their operation. Oscillator design is one such category of components that can benefit from small-signal techniques.

S-parameters are based upon matrix algebra and the linear addition of incident and reflected voltages. By definition, therefore, they are a linear technique for describing the device. The device being described is assumed invariant to the magnitude (or phase) of the incident and reflected voltages. The attempt to apply them to a device that clearly does not fall into this category is known as quasi-linear analysis. Quasi-linear techniques will be covered shortly in this chapter. Before we go on, however, we need to be clear about what we mean by a *nonlinear* circuit and its impact on component operation.

5.1 Nonlinear concepts

Let us assume we characterize a system by a black box as shown in Figure 5.1, with input signal $x(t)$ and corresponding output signal $y(t)$. Engineers will typically say the system is nonlinear if the output power is a nonlinear function of the input power.

However, we should be more precise. Strictly speaking, if x and y refer to voltage or current at the system terminals, then the system is nonlinear if

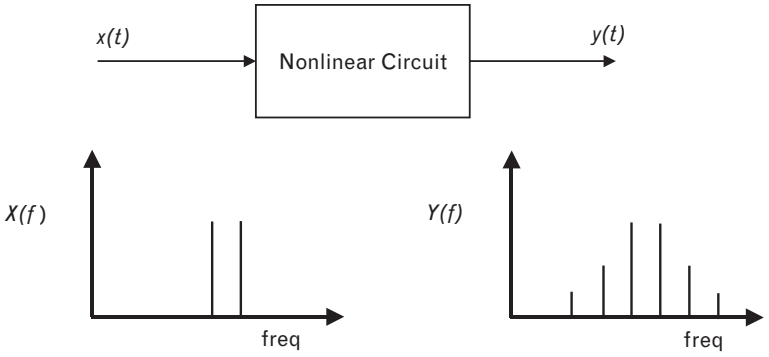


FIGURE 5.1
A characteristic description of a nonlinear circuit.

the output voltage or current is a nonlinear function of the input voltage or current—to be more descriptive, if linear superposition does not apply. Such a definition immediately rules out the application of impedance or admittance parameters to characterize the system, since, matrix algebra as then implied requires the formation of linear combinations of currents or voltages at the selected ports.

We can observe this effect by performing a power sweep using a network analyzer. The output power versus frequency at 10-mW input power level for an amplifier is as shown in Figure 5.2(a), where we assume that 10 mW is a low input power for this amplifier. As we increase the input power to 50 mW, the observed output power increases to that shown in Figure 5.2(b). In this case, the output power has not increased by the same proportion, and the device is labeled "nonlinear." The gain

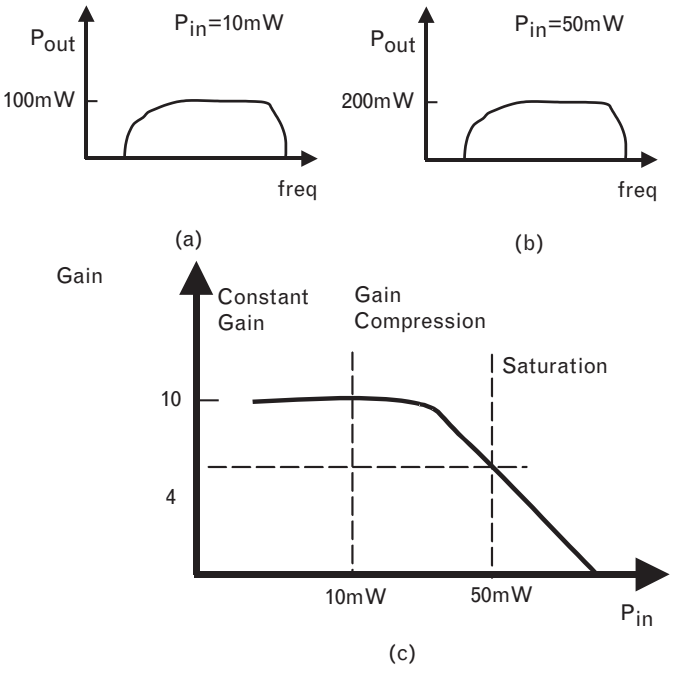


FIGURE 5.2
Power and gain characteristics of an amplifier: (a) P_{IN} = 10 mW, small signal; (b) P_{IN} = 50 mW, amplifier is saturated; and (c) amplifier gain characteristic.

plotted versus input power, shown in Figure 5.2(c), shows a region of constant gain, a region of gain compression, and a region of saturation where the output power remains fairly constant. In the region of constant gain, the device is said to be "linear" because the output power increases linearly with input power. As the device compresses, its gain begins to drop until, in the saturated region, any further increase in input power causes no additional increase in output power. In saturation the gain drops by 1 dB for each 1-dB increase in input power.

As RF engineers, we are generally more comfortable thinking in terms of the spectral content of an output signal rather than its waveshape. However, nonlinearities are usually expressed as functions of the instantaneous values of the voltages and currents related to device terminals. For instance, if the black box in Figure 5.1 is a nonlinear system, it could be characterized by a transfer characteristic relating the output to the input by

$$y(t) = \mathbf{N}[x(t)] \tag{5.1}$$

where $\mathbf{N}$ is a nonlinear function. Simple expansion of $\mathbf{N}$ as a Taylor's series will result in terms such as $x(t)^2$. For periodic inputs, such as would be indicated by the input spectrum in Figure 5.1, a consequence of this nonlinear transfer characteristic is that new frequency components will be generated, as illustrated at the output in Figure 5.1. The phase of the output signal will also be affected, depending on magnitude and phase of the input signal. We have already discussed this as it applies to intermodulation distortion in a radio.

Generally, for any nonlinear system, the consequences of a nonlinearity acting on a periodic input signal will be that the output signal has new frequency components generated. This extends, as we shall see, even to new dc components that flow in the bias network. Effects such as cross-modulation (the transfer of modulation from one carrier to another) and intermodulation (the generation of new carriers in adjacent channels) also arise. Such effects are commonly referred to as *distortion*. We will model distortion principally by looking at the amplitude and phase relationship of an output signal as a function of the amplitude and phase of the input signal.

Some caution is needed in terminology at this point. Consider the system represented by $h(t)$ shown in Figure 5.3, in which a periodic square wave pulse passes through the system and is transformed into a signal as shown. If such signals were monitored on an oscilloscope, the observer

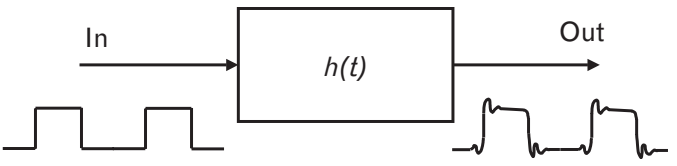


FIGURE 5.3
A periodic pulse train passing through a linear system and experiencing distortion.

would unquestionably refer to the output as *distorted*. However, if the input signal were to be doubled in amplitude and the output signal faithfully followed by doubling in amplitude as well, the system is still clearly linear according to our definition above. Of course, a simple filter consisting of inductors and capacitors could be used to implement $h(t)$. This so-called distortion, which is linear distortion, results from a change in the relative magnitude of each frequency component of the signal as imposed by the frequency response of the filter. It also results from the different frequency components of the pulse taking different times to traverse the filter. However, this is not a nonlinear effect because superposition still applies and no new frequency components are generated. Linear distortion has no amplitude or phase dependence on the input signal, and is instead characterized by the small-signal gain and its phase (or group delay) over frequency.

5.1.1 Some nonlinear phenomena

In Section 3.2.4 of Volume I, we consider power-in power-out relationships of the form shown in Figure 5.4.

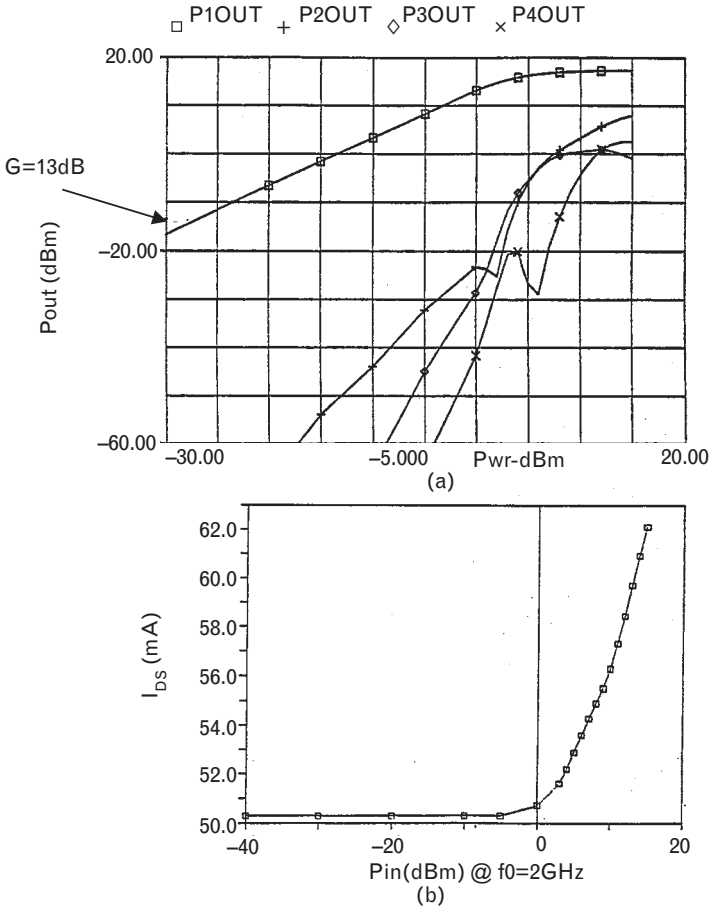


FIGURE 5.4
The dependence of (a) output power and (b) dc bias current on the input power for a typical power amplifier.

The rise of the harmonics was described by modeling the transistor output current as a power series as a function of input voltage. The higher-order terms of the power series gave rise to components containing the second and higher harmonics of the fundamental, and for a narrow power range over which the power series was valid, we showed that the second harmonic would rise twice as fast with input power as the fundamental. However, Figure 5.4(a) shows that the amplifier passes through a *sweet spot*, where the even harmonics actually fall for a brief interval over which the input power is increased. This cannot be explained simply by expanding the transconductance as a truncated power series, as we do in Volume I; instead, the model needs to be expanded to account for the saturation and turnoff effects of the overdriven transistor. Similarly, Figure 5.4(b) shows the rise of the bias current as the amplifier begins to saturate, due to a rectification effect that will require more complex models to describe its behavior. Progressively as the transistor input power is increased and the transistor moves into saturation, more complex models are required to describe effects such as harmonic generation and change in bias point.

Figure 5.5 shows a plot of the output power against input power for a typical medium-power transistor. The axes plot power in milliwatts on a logarithmic scale (which is linear if power is expressed in decibels). The

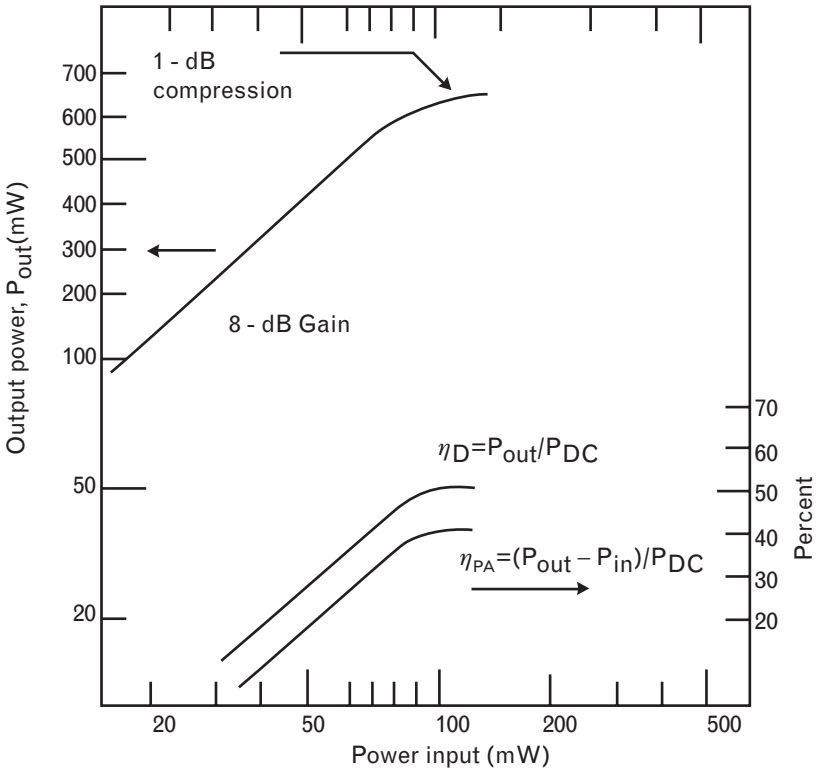


FIGURE 5.5
The dependence of total output power, efficiency, and power-added efficiency with input power for a typical transistor.

output power is 120 mW when the input power is 20 mW, so the small-signal gain is 6, or about 8 dB. When the gain has dropped to 7 dB, the amplifier is said to be in 1-dB compression, and this occurs at around 120 mW of input power.

The drain efficiency is defined as the ratio of RF output power to dc input power, so rises in roughly the same proportion as the fundamental output power if the dc power is constant. The degree of correlation depends on the relative contributions of the harmonics to the total RF output power, and the degree by which the dc input power changes as the device enters saturation. The efficiency is an important parameter in power amplifier design, and in later sections of this chapter we will explore ways to maximize it.

The power-added efficiency is defined by

$$\eta_{PA} = \frac{\left(P_{OUT} - P_{IN}\right)}{P_{DC}} \tag{5.2}$$

and measures the incremental RF power added by the device, comparing the output power to the level of input power needed to achieve it. This measure of efficiency depends on the gain of the device G since $P_{OUT} = GP_{IN}$. It is a useful performance measure in amplifier design because it tells us the relative contribution and cost made by the device to enhancing power levels. The power-added efficiency always has the same shape: concave down, because with no input power it is zero, and at very high power levels the input power can exceed the saturated output power so it becomes negative. Therefore, the power-added efficiency must pass through a maximum value.

Figure 5.6 shows the relationship between the output and input power in more detail for another device. The 1-dB compressed output power is about 24 dBm and occurs at about 19 dBm input power. When the input power (P_{IN} in milliwatts) is subtracted from the output power (P_{OUT} in milliwatts), the difference curve $P_{OUT} - P_{IN}$ can be constructed as shown. If the dc power is constant, then this curve mirrors the power-added efficiency, except for scale. The point of maximum $P_{OUT} - P_{IN}$ in this case, and in fact generally, occurs around the 1-dB compression point of the matched device. This is because the 1-dB compression point marks the boundary around which P_{OUT} is close to its maximum but the gain is still sufficiently high that P_{IN} can remain reasonably small. Furthermore, the peak value of $P_{OUT} - P_{IN}$ measures the maximum power that can be obtained if this device were to be used in an oscillator, since then P_{IN} must be subtracted from the output power to sustain the oscillation. The remainder $P_{OUT} - P_{IN}$ is left over for the oscillator output. This is another fundamental result: the maximum output power of an oscillator is determined a priori by the device and can never exceed the peak value of $P_{OUT} - P_{IN}$ of a matched device, which

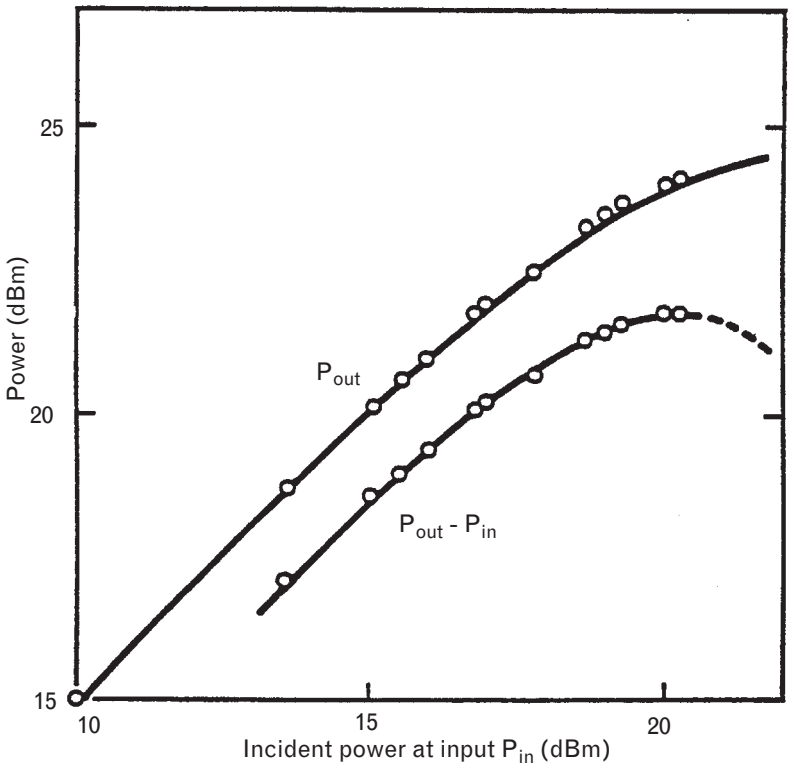


FIGURE 5.6
Magnified plot of the output power versus the incident input power for a transistor.

generally occurs around the 1-dB compression point. The function of the oscillator circuit is to feed back the right proportion of the output power, in the right phase sense, to drive the device at this operating point.

This last case illustrates an important example of nonlinear design philosophy: the device is embedded in a linear circuit, and the function of the circuit is to impose the proper boundary conditions on the device behavior. For example, the function of an oscillator circuit is to drive the device close to its 1-dB compression point by feeding back the right proportion of output power. Another example of imposing boundary conditions is the amplifier load line. In the next section we will see that the function of the output circuit in that case is to ensure the device sees the appropriate output resistance. This constrains the device output voltage and output current waveform in such a way as to deliver maximum power to the load.

5.2 Quasi-linear power amplifier design

Earlier chapters of this book have covered in great detail the principles of small-signal design, in which the device is assumed invariant to whatever circuit it is embedded in. Small-signal objectives are typically to design the output matching network for complex conjugate match, or to flatten the

gain over frequency by introducing loss at the lower frequencies. The input network is then designed for the appropriate criteria, perhaps for optimum gain or noise match.

With design at large-signal levels to achieve significant output power, there is little difference to the design of the input network, apart from feedback effects through the s_{12} of the device itself. However, the output network must now be designed to extract maximum power out of the device. The output network is critical to high-power designs, because the output is where the voltage and current swings of the device are high, where these need to swing in phase, and in such a way as to minimize distortion and maximize efficiency.

The most useful tool for analyzing these signal swings at the output is the load line imposed on the device. The load line is rich in information, for it is centered around the bias point of the device—its length indicates the level of signal swing, while its slope the load impedance. The load line traverses different regions of the device I-V curves, and close examination can reveal the device operating regions, its instantaneous output power, where distortion arises, and even its efficiency. In all of the descriptions below, we will alternate between using the notation either for a (depletion-mode) FET such as a MESFET, or a bipolar device; and unless otherwise indicated the analysis applies equally to either device.

5.2.1 The amplifier load line

Consider the typical amplifier circuit of Figure 5.7, in which a MESFET is biased through an RF choke with a quiescent drain bias voltage of V_{DD} and quiescent drain bias current I_Q. The output blocking capacitor will charge to a steady-state value V_{DD} whenever the output voltage V_O swings low, and assuming it is a large enough dc blocking capacitor, will remain charged at that value throughout the entire RF cycle. We may then write

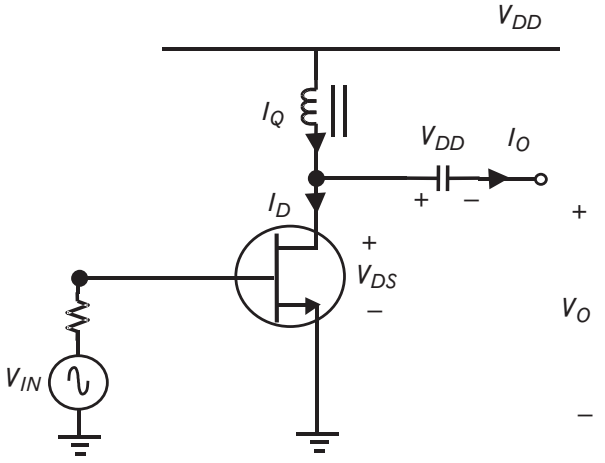


FIGURE 5.7
A device embedded in a simple circuit for calculating its load line.

$$V_{DS} = V_{DD} + V_O \tag{5.3}$$

using capital letters to indicate total voltage and current (dc included). If we take incremental quantities instead, we see that $v_{DS} = v_O$, that is, maximizing the drain voltage swing also maximizes the load voltage swing. The current through the RF choke will also be constant, and we can write

$$I_Q = I_D + I_O \tag{5.4}$$

Again, if we take incremental quantities $i_D = -i_O$, so maximizing the drain current also maximizes the current in the load.

We now impose a constraint on the device output voltage and current by introducing a load impedance $Z_L = V_O / I_O$, so that we now require

$$\begin{aligned} I_D &= I_Q - I_O \\ &= I_Q - \frac{V_{DS} - V_{DD}}{Z_L} \end{aligned} \tag{5.5}$$

This last equation is a fundamental equation describing how the transistor output current I_D changes with voltage V_{DS}. The circuit now imposes a boundary condition on the MESFET drain current and voltage, forcing the current and voltage to lie along the trajectory defined by (5.5). This trajectory is centered around the bias point of the device, for when $V_{DS} = V_{DD}$, then $I_D = I_Q$. The trajectory is known as the load line, and if Z_L is a real resistance of value R_L, the slope of the load line will equal $-1/R_L$ and its length will be determined by the amplitudes of the current and voltage swings at the drain, since in the above circuit $Z_L = V_O/I_O = v_O/i_O = -v_{DS}/i_{DS}$.

It is normal to superimpose the load line on the I-V curves of the device itself, as shown for an FET in Figure 5.8. In the first instance, because the reactive output parasitics of the device are invisible at dc, the load line is essentially a dc relationship, in which the voltage at the intrinsic device terminals (i.e., across the output current source) is now constrained to obey (5.5). However, there are further constraints on the load line trajectory imposed by the device itself, as follows:

- The minimum device voltage is the knee voltage, V_{SAT}, close to zero.
- The maximum device voltage is limited by its breakdown voltage.
- The minimum drain conduction current into the device is zero and cannot go negative.
- The maximum drain conduction current into the device is I_{MAX}.

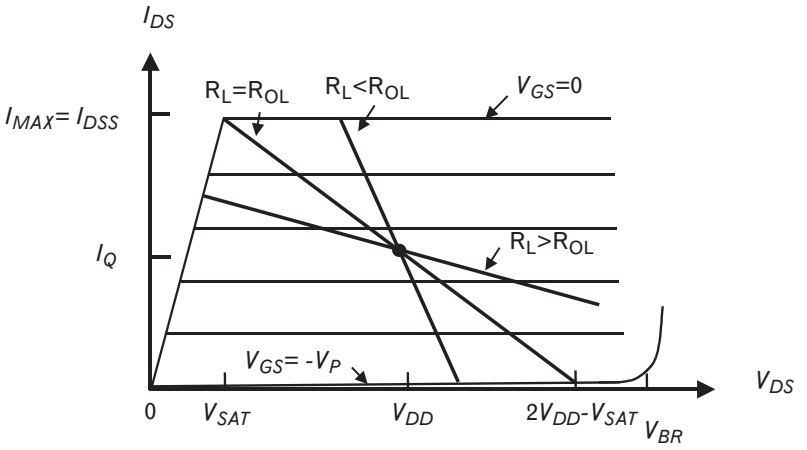


FIGURE 5.8
Output load line for a device plotted with three different values of load resistor.

Since these are implicit to the I-V curves themselves, the load line is plotted according to (5.5) but its trajectory is ultimately limited by these device constraints. In passing, we should remember that exceeding these constraints can be catastrophic for the device. For instance, operating a device into a high VSWR load could exceed the breakdown voltage on the drain or the maximum current into the device, depending on the phase of the load. Although the expected load line will be plotted so it avoids these excess conditions, a faulty output connection or unusual load on the antenna could potentially cause the actual load line to be quite different and lead to failure of the device unless protection circuitry is built in.

The curves of Figure 5.8 correspond to three different values of load resistor. These curves all pass through the device bias point. In Figure 5.8, this bias point corresponds to a point (V_{DD},I_Q) and is achieved when the gate voltage is approximately $-|V_P/2|$, midway between 0V (when the device is switched full on) and the pinch-off voltage $-|V_P|$ (when the device is turned off). The quiescent current that results is then approximately $I_{MAX}/2$, where $I_{MAX} \approx I_{DSS}$ for a MESFET, corresponding to the current when the gate voltage is zero.

Consider now the case when $R_L = R_{OL}$. If we assume a sinusoidal gate voltage, then as the gate voltage swings positive from $-|V_P/2|$ to zero, the drain current can rise from $I_{MAX}/2$ towards I_{MAX}. The gate voltage can, in fact, instantaneously swing slightly positive to the point of the gate-source diode entering forward conduction, and although this increases the maximum current swing I_{MAX} beyond I_{DSS}, it will increase the distortion and reduce the lifetime of the device. The output voltage is constrained to lie along the load line given by (5.5), and as shown in the Figure 5.8, will fall from V_{DD} to the knee of the curve V_{SAT}. Assuming sinusoidal output voltage and current across the load can be maintained by a circuit with reasonably high Q at the collector, and neglecting harmonics, the resulting output current and voltage are sinusoidal with zero-to-peak (or *peak*, for short) amplitudes $I_{MAX}/2$ and $V_{DD} - V_{SAT}$, respectively. When the gate voltage swings in the opposite

direction down to pinch-off, the drain current falls from $I_{MAX}/2$ to zero, and the output voltage rises from V_{DD} to $2V_{DD} - V_{SAT}$. The RF choke permits the maximum collector voltage to swing symmetrically to almost twice the rail voltage, unlike the low-frequency dc-coupled audio amplifier where the choke is replaced by a collector resistor. Here we have constructed a typical class-A amplifier, in which the device is always conducting and the load line remains within the active region of the device.

The output resistance R_{OL} can be calculated from the slope of the load line derived from its endpoints, giving

$$R_{OL} = \frac{2\left(V_{DD} - V_{SAT}\right)}{I_{MAX}} \tag{5.6}$$

This is an important equation, which states that the optimum load resistance for a device is a function of the device itself and the bias point. If V_{SAT} is close to zero, then if the bias voltage is doubled, the output resistance is doubled for the same current swing.

The equation also states that for higher-power devices, those with larger maximum currents, the optimum load resistance becomes increasingly smaller. Device manufacturers can create a 1-W device by placing two 0.5-W chips in parallel. The bias voltage is unchanged, but the maximum current is thereby doubled to give double the power. Therefore, the optimum load resistance of the 1-W device is one-half that of the 0.5-W device, and can become agonizingly low as devices get bigger. It is for this reason that the bias voltages are made as high as allowable whenever possible: to maintain a reasonable level of matching impedance for R_{OL}.

The output power can be calculated from the load line if we know the magnitude of either the voltage or current swing at the output. The power into the load is

$$P_{RF} = \frac{I_{PEAK}V_{PEAK}}{2} = \frac{I_{MAX}\left(V_{DD} - V_{SAT}\right)}{4} \tag{5.7}$$

where we will use the subscript "*PEAK*" to represent a sinusoidal signal with the indicated peak amplitude measured between its average (zero) and its peak value. This equation shows why the curve $R_L = R_{OL}$ in Figure 5.8 corresponds to an optimum (power) load. Maximum power is delivered whenever we can obtain maximum output current swing, and maximum output voltage swing, in phase. Equations (5.3) and (5.4) show this is achieved when the device current and voltage swings are also at their maximum. But the device constraints are such that the maximum current swing is limited to a peak-to-peak value of I_{MAX} and a peak-to-peak voltage swing that extends from the knee of the curve to breakdown V_{BR}. In this case, the maximum power a device is capable of producing is

$$P_{MAX} = \frac{I_{MAX}\left(V_{BR} - V_{SAT}\right)}{8} \tag{5.8}$$

where all the quantities in (5.8) are intrinsic to the device itself. This power will be obtained when the device bias voltage is set at $V_{DD} = (V_{BR}+V_{SAT})/2$, so the drain voltage can swing fully and symmetrically between the knee and the onset of avalanche in the reverse-biased gate-drain diode. We have assumed here that thermal or bias limitations do not constrain this choice of current and voltage swings.

In most instances, however, the bias voltage is not necessarily a free parameter, and the battery voltage will predetermine the value of V_{DD}. In this case, the maximum voltage swing that can be achieved is a zero-to-peak voltage of $(V_{DD} - V_{SAT})$, and the zero-to-peak current swing can be $I_{MAX}/2$. To achieve these swings simultaneously and in phase at the terminals of the device current source requires the load resistance R_{OL} given by the optimum value in (5.6). It is important to specify the reference planes for the value of optimum load resistance presented to the device, because it is the *intrinsic* device voltage and current that must be in phase. The effects of any device parasitic reactance and load mismatch need to be nulled out by the output matching network between the device and the actual load resistor. This implies that the current and voltage at the external drain terminals of the device may, in fact, not be in phase. At the actual load resistor itself, of course, the current and voltage will again be back in phase because this is the constraint a resistor imposes on the relationship between voltage and current. By absorbing the device parasitics in this way, the restriction we had originally placed on this being a dc analysis can be removed.

Two other load lines are plotted in Figure 5.8. The first of these, with $R_L < R_{OL}$ will have a steeper slope than the optimum because of (5.5). In this case, as the gate voltage swings between zero and pinch-off, the drain current swings between I_{MAX} and zero. However, because the load resistor is too small (compared to the optimum), the output voltage swing is not as large as before. This circuit is said to be *current limited* because the maximum current swing limits the output power.

The second case in Figure 5.8 is for $R_L > R_{OL}$, and in this case the voltage swing is the maximum attainable for the device, with the drain voltage dropping to V_{SAT} when the gate voltage rises to zero, and the drain voltage rising to $2V_{DD} - V_{SAT}$ when the device switches off. In this case the current swing is not as large as before, because the load resistor is too large and the output current cannot swing between its potential peak values. The circuit is said to be *voltage limited* in this case because the bias voltage limits the voltage swing that can be attained for the chosen load resistor.

The load line is also useful for determining where the distortion arises in a device. Consider Figure 5.9, which shows a small signal amplifier and its load line plotted on the I-V curves of the device. As the input base

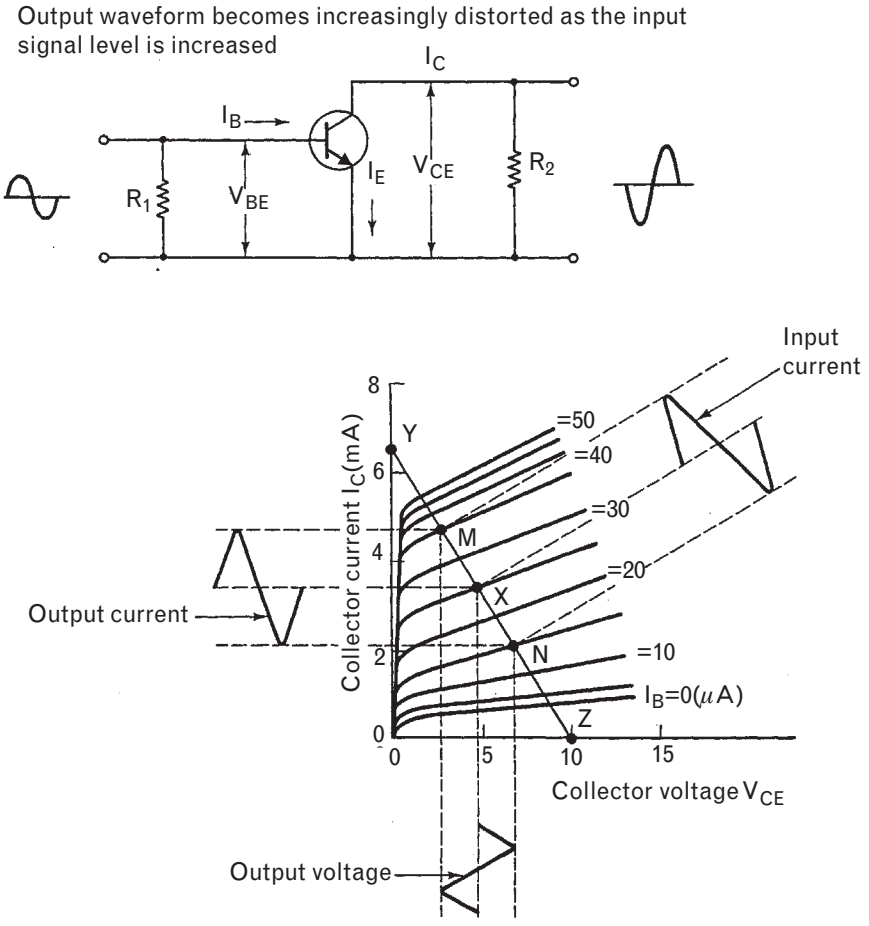


FIGURE 5.9
Use of the load line for a transistor amplifier.

current swings sinusoidally around its bias point, between 15 and 35 μA, the output current and voltage swings will also be sinusoidal because the spacing between the current curves is fairly constant with incremental base current. However, as the input current swing increases further, the drive moves into regions where the device begins to cut off, and goes into saturation. As a result, the output current and voltage begin to flatten at the peaks of the sinusoids. We will see later that this results in a rapidly increasing third harmonic content at the output. Ultimately, as the device is overdriven, the fundamental device limits assert themselves: the output device voltage is clipped at the knee voltage when it swings low, and the output device current cannot swing negative when the device switches off during the other half cycle. With the bias voltage, these limits determine the saturated output power that results.

So far, we have only considered resistive load lines. But it is possible that the load impedance can contain reactance, and Z_L in (5.5) will not be a real number. Consider, for instance, the case when Z_L is a capacitor; the

RF load current will then lead the load voltage by 90° and the RF drain current, which is the opposite of the RF load current, will thus lag it by 90°. Thus in (5.5), when the drain voltage is close to zero and minimum, the drain current has yet to fall to its minimum. The load line becomes elliptical about the bias point, with the current and voltage out of phase. The ratio of the two axes is proportional to the reactance of the load capacitance. We will see this effect later in some of the power amplifier examples, when the impedance presented at the measurement terminals contains a reactive component that is not nulled out by the matching network.

We have also assumed that at dc the collector bias voltage is equal to the supply voltage. The presence of any series collector resistor R_C in the bias network will change the slope of the load line, since, seen from the collector, it appears in parallel with the load resistor. In fact, the load line measured at dc, which was previously vertical because the dc impedance of the RF choke is a short circuit, now takes on the slope $-1/R_C$. It changes to slope $-(1/R_C + 1/R_L)$ as the blocking capacitor at the output becomes a short circuit at high frequencies.

The presence of harmonics in the waveform will also change the appearance of a load line, although if the load is purely resistive at the harmonic frequencies as well as at the fundamental, the presence of harmonics is not obvious from the load line itself. Rather, as the voltage and current traverse the load line, their movement is not monotonic along it; the harmonics can cause the direction of the trajectory to change or to stop altogether if the voltage and current are constant during part of the cycle (as, for instance, with square wave voltages and currents). However, it is rare (and usually undesirable) that matching circuits will terminate the harmonic components resistively, so that the reactance of the harmonic impedances is usually obvious as the load line will open up or bend, often in peculiar ways.

As a final note on the calculation of output power from a device, we should note that expressions such as (5.7) or (5.8) are invariant with frequency. In other words, a device that can deliver a 1-dB compressed output power of 20 dBm at 500 MHz in theory has the same capabilities at 1 GHz. However, whether that power can literally be extracted from the device itself is another matter. We have made these calculations at the internal current terminals of the device, intrinsic to the device itself. Between that current source and the external terminals of the device is the shunt capacitance of the device itself and the package, and a lead with series inductance. These parasitic elements form a perfect lowpass filter, although normally their 3-dB frequency will lie beyond the usual frequency of device operation. Beyond that, of course, it will become increasingly difficult to extract the power predicted by these equations, as the achievable power will roll off according to the frequency response of that internal lowpass filter.

5.2.1.1 The load lines for maximum power and maximum gains

We have seen that the load line imposes a locus on the output current and output voltage of the device. From the infinite combinations of output current and voltage accessible in the device output I-V space, only those satisfying the constraint imposed by the load impedance can be reached.

Consider the I-V curves of Figure 5.10. We saw in Chapter 3 that a device with these output curves can be modeled by an output current source in shunt with an output resistor r_o, where at low enough frequencies, r_o is the slope of the current curves in their "flat" region. Device avalanche at the breakdown voltage is apparent in this example, but this is not essential in the consideration that follows.

Our analysis so far has assumed that the I-V curves from dc are still valid in the RF operating region for the device. This is not strictly true for a number of FETs, because the output resistor r_o can decrease significantly with frequency. Some MESFET I-V curves can also show a fictitious region of apparent negative resistance, where the drain current apparently decreases with increasing drain voltage. This effect disappears if the I-V curves are measured at a high enough pulse rate, to avoid thermal effects and the influence of traps under the gate, on the measurement. Nonetheless, if the principle use of the load line is to determine the optimum load resistor for maximum output power, these are set by the fundamental device limits and such issues are secondary. If we now make the further assumption that the device is unilateral, then the slope of the load line for maximum gain will be that corresponding to a load impedance termination with reflection coefficient s_{22}^*. But at low frequencies the output reflection coefficient s_{22} of the device will just be that corresponding to r_o, so its conjugate will also be r_o and the slope of the load line for maximum small-signal gain will be $-1/r_o$. This is just the negative of the slope of the I-V curves and is evident in Figure 5.10.

The load line for maximum output power is also shown in Figure 5.10. Its slope, equal to $-1/R_{OL}$, is set by maximizing the drain current and voltage swings for the given bias voltage. In this case, the maximum current is

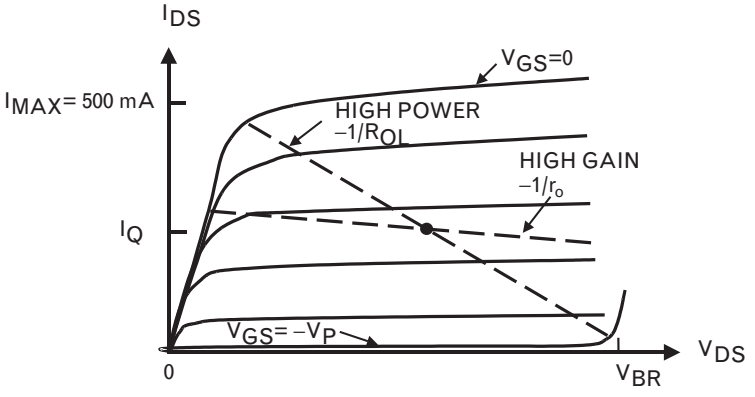


FIGURE 5.10
Load lines of a MESFET, for either maximum gain or maximum power.

500 mA corresponding to I_{DSS}, and the maximum voltage swing is between the knee and the onset of breakdown. It is clear in this example that the slope of the load line for maximum power is considerably steeper than that for maximum small-signal gain, and that here, $R_{OL} < r_O$. What is the difference then between the two lines? Figure 5.11 helps clarify.

At small signal levels, the load resistor r_o must give more output power than R_{OL}, because the former is optimized for maximum small signal gain. The output current source appears to have a source impedance of r_o at small-signal levels, so for maximum power transfer at small signals, the load resistor will also be r_o. However, the picture is quite different at large-signal swings, where the ultimate limits on current and voltage across the output current source limit the output power. R_{OL}, by constraining the voltage and current in such a way as to define a trajectory that captures the maximum possible simultaneous voltage swing and current swing, forces the current source to assume an internal impedance of $V_O/I_O = -R_{OL}$, if current is defined as coming out of the device. Thus at high power levels, the output power (and consequently the large-signal gain) with R_{OL} is higher than with r_o. The saturated output power with r_o is not as high, because (in this example) the output current swing is not as large as it could be. It is important to recognize that for the small-signal design, further increase in drive will not increase the saturated output power because the device is voltage limited. This is fairly typical when a small-signal amplifier is driven into saturation, because the small-signal output resistance is typically higher than the optimum load resistance. As a consequence, the saturated output power and even the 1-dB compressed output power of a small-signal amplifier driven with large signals will be less than the maximum achievable for the device employed.

5.2.2 Load pull methods

The load resistor presented to the intrinsic device terminals is obviously fundamental to constraining the device output voltage and current, and therefore in determining the output power. The optimum load is always a resistor at the internal current source of the device, and once this is

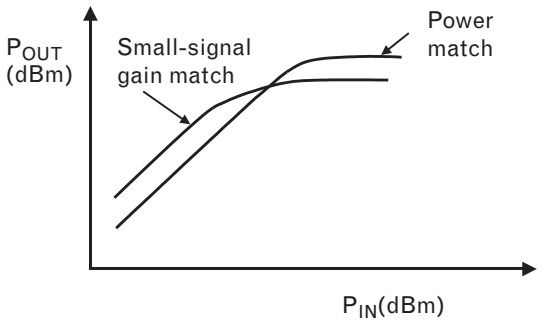


FIGURE 5.11
Plot of output power versus input power for a device matched either for maximum small-signal gain, or for maximum output power.

transformed through the device parasitics, becomes the optimum load impedance Z_{OL} that must be placed at the external device terminals. There is just one impedance that gives maximum output power.

Just as we did for gain or noise, it is possible to construct the loci of impedances that give less power than the maximum. A load pull measurement test set can be used to vary the impedance Z_L seen at the device terminals and to measure the resulting output power.

The test set can be either active or passive. With a passive measurement system, seen in Figure 5.12, any impedance Z_L within the entire Smith chart can be created using double stub tuners. The positions of the stubs along the transmission line, and the distance between them is varied to change the impedance. The device under test is driven at its input by a fixed large-signal voltage V_1^+ that is matched to the device through an input tuner. This tuner is readjusted at each measurement point (as the output is changed) to maintain conjugate match at the input and thus constant input power. The output tuner is changed to find the locus of Z_L that keeps constant fundamental output power. A series of contours result for different defined output power levels.

With an active measurement system, the output is loaded not with a variable impedance but with a second variable source V_2^+ whose amplitude and phase are varied. The impedance at the output of the device is then

$$Z_L = \frac{V_2}{I_2} = \frac{V_2^+ + V_2^-}{V_2^+ - V_2^-}$$

and because V_2^- can be varied independently of V_2^+ by means of V_1^+, any arbitrary impedance can be created at the output.

The resulting loci are called load pull contours, and each contour represents the maximum output power achievable with a given load impedance on the device. They are useful for determining the actual impedance a device should see when used in an amplifier, and provide an alternative to (5.6) and the device package model for calculating equivalent results.

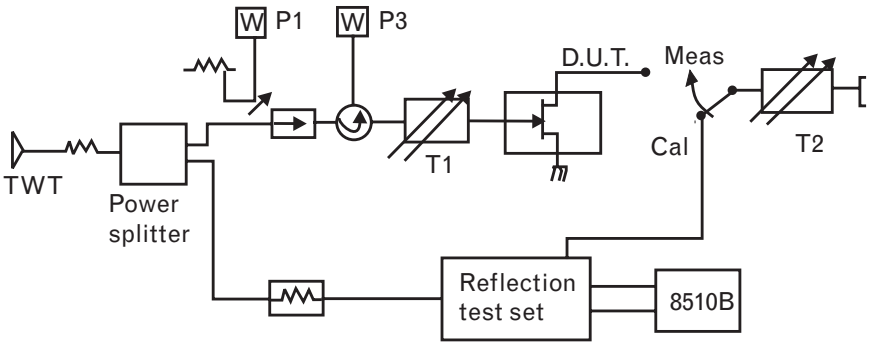


FIGURE 5.12
A passive load pull measurement system.

Figure 5.13 shows an example of load pull contours for a typical MESFET. The point of maximum output power is in the middle of the contours, and is a single impedance, corresponding to Z_{OL}. The first contour is drawn for a power level 0.5 dB less than the maximum output power. Analogous to the circles of constant operating gain for small-signal operation, there are an infinite number of impedances that can achieve this output power level (or this level of constant large-signal operating gain). Unlike the operating gain circles, however, the load-pull contours are not circular. We will derive approximate expressions for them shortly.

As the power is reduced further, to 2 dB less output power than the maximum, the measured load pull contour for the device in the figure is not closed. There is a range of impedances missing from the contour near the edge of the Smith chart, possibly because either:

- The device becomes unstable when presented with these output impedances, and starts to oscillate;
- The device operation becomes unsafe when presented with these output impedances—for instance, if the drain current exceeds some limit;
- The tuner is unable to synthesize impedances (near the edge of the Smith chart) which have low loss. This would be a problem indeed,

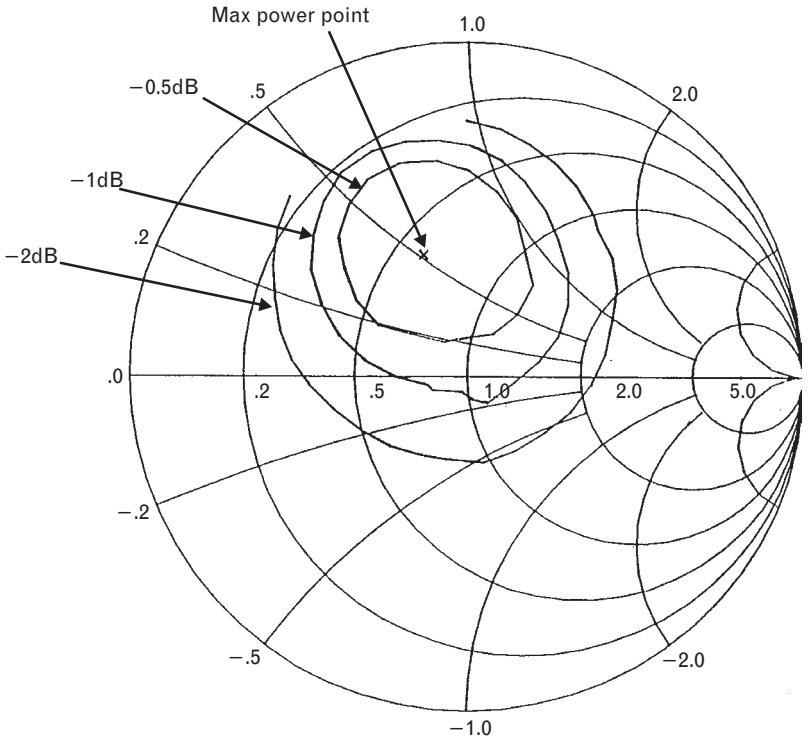


FIGURE 5.13
The loci of load pull contours for a power MESFET. The optimum power point and contours for 0.5, 1, and 2 dB less output power are shown.

for optimum lead impedances are typically very low, of the order of several ohms, for power transistors. For instance, a 1-W device biased at 3-V drawing 0.7A of quiescent current will have an optimum load impedance of just over 4Ω, using (5.7) and (5.6).

Recently, a number of CAD simulators have introduced the capability to perform load pull simulations on an active device, by analyzing a number of different load impedances across the Smith chart and interpolating the resulting output power to find impedances with equal output power. However, this serves to illustrate a difficulty with load pull measurements, and that is that the terminations at the harmonic frequencies are usually quite arbitrary and rarely, if ever, measured. Although the load pull tuners can accurately fabricate a given load impedance, this is at a known fundamental frequency, and the harmonic impedances that result from the particular tuning settings are usually not characterized. For example, harmonic terminations can make a large difference to the efficiency of a device, so load pull measurements need to be very carefully interpreted if contours of constant efficiency are measured instead of constant output power. For the fundamental output power, the impact of changing the harmonic termination between an open and a short circuit is rarely more than 1 dB, so the error in neglecting harmonics is relatively small if the device is not heavily saturated.

5.2.2.1 Predicting the load pull power contours

It is possible to predict the location of the output power contours using the simple quasi-linear theory we have already used [1]. Although most simulators will automatically calculate these, the analysis is still useful since it illustrates the mechanisms that limit the output power. For this analysis, which is illustrated with an FET but in principle is similar for a bipolar device, we will assume as above class-A operation and an ideal device for which the saturation voltage is approximately zero.

For maximum linear power at the onset of compression, the RF load must be resistive and will appear at the intrinsic device terminals from (5.6) as

$$R_{OL} = \frac{2V_{DD}}{I_{MAX}} \tag{5.9}$$

$I_{MAX} = 2I_Q$ and will equal I_{DSS} only if the bias current I_Q allows it. This equation defines the output power contour for the case of maximum output power sustainable by the device under the given bias conditions. The contour collapses to a single point.

We next consider output powers that are less than this theoretical maximum. There are two cases:

- *Case I*, where $|Z_L| < R_{OL}$. This is the current limited case because the output power is limited by the maximum current swing.
- *Case II*, where $|Z_L| > R_{OL}$. This is the dual of the above case and is voltage limited because the output power is limited by the maximum voltage swing.

In each case, the device input power is assumed sufficient to still drive the device to the point of either maximum current or maximum voltage, so that the maximum power for the given load impedance is measured. The bilateral effects of the load on the source impedance ("source-pulling") are accounted for by retuning the input to keep the input matched and the input power constant. The load line and waveforms for the two cases are given in Figure 5.14.

In case I, just prior to compression, the device is driven sufficiently hard so that the current swings the entire available swing (i.e., from zero to I_{MAX}). Because the magnitude of the resistance is too low, the voltage never attains its full possible swing.

In case II, just prior to compression, the device is driven sufficiently hard so that the voltage swings the entire available swing [i.e., from close to zero (or more precisely, V_{SAT}) to approximately twice the supply voltage ($2V_{DD} - V_{SAT}$)]. Because the magnitude of the resistance is too large, the current never attains its full possible swing.

For case I, consider the load impedance as a series impedance $Z_L = R_L + jX_L$. The maximum linear power is

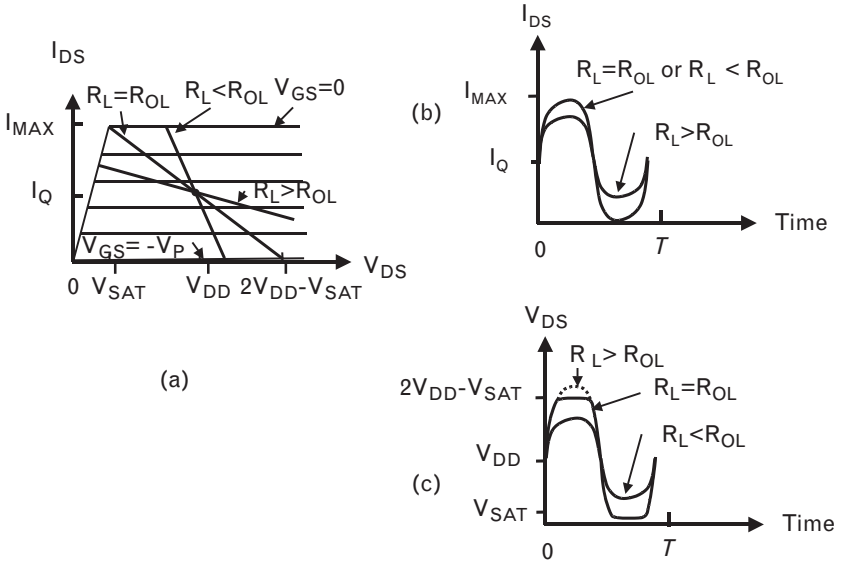


FIGURE 5.14
Derivation of maximum power available from a device into a give load impedance: (a) the load line; (b) the drain current; and (c) the drain voltage.

$$P_L = \frac{1}{2} I_{PEAK}{}^2 R_L = \frac{1}{2}\left[\frac{I_{MAX}}{2}\right]^2 R_L \tag{5.10}$$

since we know in this case we have the full current swing. Of course, if $R_L = R_{OL}$ then $P_L = P_{OPT}$ and we can write an identical expression. Normalizing,

$$\frac{P_L}{P_{OPT}} = \frac{R_L}{R_{OL}} \tag{5.11}$$

The peak-to-peak drain voltage can then be calculated and is given by the (known) peak-to-peak current times the impedance

$$|V_L| = I_{MAX}\sqrt{R_L^2 + X_L^2} \tag{5.12}$$

Substituting for I_{MAX} from (5.9), we obtain

$$|V_L| = \frac{2V_{DD}}{R_{OL}}\sqrt{R_L^2 + X_L^2} \tag{5.13}$$

Since case I is current limited, the voltage swing never attains the full swing possible, $2V_{DD}$ (if we assume V_{SAT} can be neglected). In order to keep $|V_L|$ less than this, in (5.13) we must have

$$|X_L|^2 \leq \left(R_{OL}{}^2 - R_L^2\right) \tag{5.14}$$

where the equality sign occurs when $|V_L|$ just attains the full $2V_{DD}$ swing.

For case II, a similar analysis applies, except we use the dual argument. We consider the load as a parallel admittance $Y_L = G_L + jB_L$ and since the voltage swing is known to equal $2V_{DD}$ just prior to compression, the maximum linear power is

$$P_L = \frac{1}{2}\frac{V_{PEAK}{}^2}{R_L} = \frac{1}{2}\left[V_{DD}\right]^2 G_L \tag{5.15}$$

Again, we may normalize to the optimum power, which occurs when $G_L = G_{OL}$ and for which a similar expression for optimum load power may be written, so that

$$\frac{P_L}{P_{OPT}} = \frac{G_L}{G_{OL}} \tag{5.16}$$

Since case II is voltage limited, the current never attains its full output swing, and so the susceptance must now be kept within the limits given by the dual of (5.14),

$$|B_L|^2 \leq \left(G_{OL}{}^2 - G_L{}^2\right) \tag{5.17}$$

Essentially, in both cases, it is the real part of the impedance or admittance that must remain constant if the output power remains constant, and if either the current is fixed at its limit or the voltage is fixed at its limit. Varying the reactance in case I and the susceptance in case II will vary the magnitude of the voltage or current, respectively, as well as the phase angle between the voltage and current, without affecting the output power. Thus, in both cases, the real part of the product of voltage and current, which is just the output power, remains constant, while the imaginary part, the reactive power, increases up to the point where the other voltage or current hits its limit.

This then gives a process for constructing the load pull power contours under given bias conditions, as follows:

1. Determine R_{OL} from (5.9). This gives the maximum power point.
2. For a load pull contour of a given power level, use (5.11) and (5.16) to determine the resistive points on the contour of that power level.
3. Starting at the smaller resistance, the contour follows a constant resistance line on the Smith chart up to the reactance limits given by (5.14).
4. Starting at the larger resistance, the contour follows a constant conductance line on the Smith chart up to the susceptance limits given by (5.17).
5. Transform the reference plane of the contour to the external device terminals, by absorbing the effect of the device shunt output capacitance and series bond wire inductance or package into the measurement.

The last point inevitably causes confusion because it is the inverse of the standard matching process. The exercise below will help clarify these steps.

5.2.2.2 Exercise in creating the load-pull power contours

Figure 5.15 shows a simplified output model of an FET that is capable of up to 3-W linear output power. We bias the device at 3V and 750-mA quiescent current, so take $I_{MAX} = 1.5A$ (its actual current capability is higher than

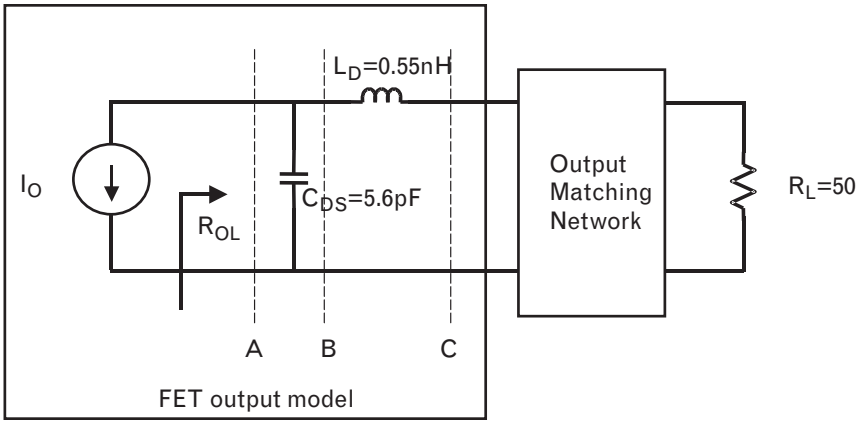


FIGURE 5.15
Transistor model for calculating the load pull power contours.

our bias circuit will allow). We want to calculate the optimum load resistor, and the –1-dB load pull power contour at the external terminals of the device, at 1,900 MHz.

The first step is to use (5.9) to calculate R_{OL} for the given conditions. With $V_{DD} = 3$ and $I_{MAX} = 1.5A$, it follows that $R_{OL} = 4\Omega$. Normalized to a 50-Ω system, this plots at 0.08 on the Smith chart in Figure 5.16(a). The maximum linear output power for these bias conditions is given by (5.10) and equals 1.125W or 30.5 dBm.

The second step is to calculate the resistive loads that can support 1 dB less power (i.e., 29.5 dBm or 890 mW). Using (5.11),

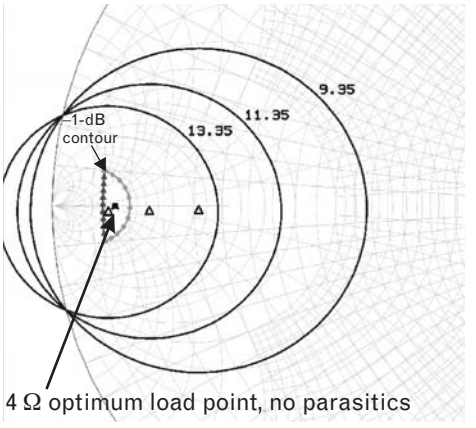
$$\frac{0.89}{1.125} = \frac{R_L}{4}$$

so that $R_L = 3.2$. Using (5.16),

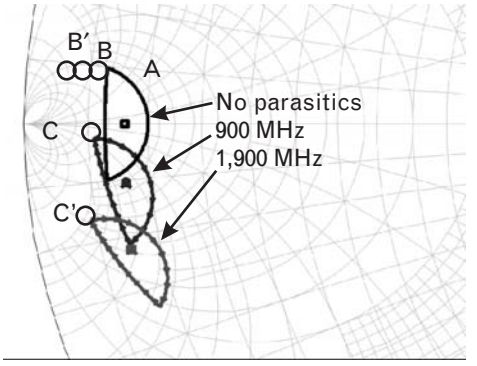
$$\frac{0.89}{1.125} = \frac{G_L}{1/4}$$

so that $R_L = 5\Omega$. These normalize to 0.064 and 0.1, respectively, on the Smith chart and form the two resistive points on the –1-dB contour plotted in Figure 5.16(a).

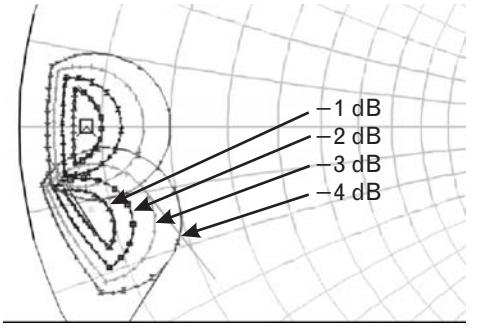
The third step is to trace an arc of constant resistance along the $R_L = 3.2\Omega$ line. The reactance limits in (5.14) could be used to define the limits of the arc, but it is probably simplest to leave the arc open-ended, and move to the fourth step, which is to trace an arc of constant conductance along the $G_L = 1/5\Omega$ constant conductance circle. Although the limits defined in (5.17) also define the endpoints of the arc, the 1-dB contour can be closed instead by drawing the arc until it intersects with the constant resistance $R_L = 3.2\Omega$ arc. At the intersection, the voltage and current swings are both at their respective limits for voltage and current limited operation simultaneously



(a)



(b)



(c)

FIGURE 5.16
(a) Smith chart showing the optimum load and the intrinsic −1-dB load pull contour, compared with circles of constant operating gain. (b) Transformation of the power contour through the parasitics of the device at 900 and 1,900 MHz. (c) Power contours from −1 to −4 dB before and after parasitic absorption at 1,900 MHz.

$$V_{PEAK} = 2V_{DD}$$
$$I_{PEAK} = \frac{I_{MAX}}{2}$$

These, of course, are the magnitudes of voltage and current swing required for maximum power from the device, but the difference now is that the voltage and current are not in phase. The phase angle between

them is such that the real part of their product will give only 890-mW output power rather than 1.125W.

The load contour shown is the range of loads that can be connected at the internal device terminals to generate the power required. To transform to the external device terminals, where any physical load must be placed, the parasitics must first be absorbed into this measurement. We start at 1,900 MHz and consider first the 5.6-pF shunt capacitor, representing the combined package and device output capacitance. Assume that the output matching network in Figure 5.15 correctly transforms the external 50-Ω load to the optimum load resistance measured at the intrinsic terminals of the device (4Ω). This optimum load and the load pull contours are the relevant termination impedances for the device (i.e., the impedance looking to the right of the figure out of the device). As we move to reference plane B, we see less shunt capacitance looking out of the device, since at reference plane A, the first element in the load was the shunt capacitor, and at reference plane B it is no longer part of the load impedance. Thus, to obtain the required load at B, the susceptance of the capacitance needs to be subtracted from the measurement at A. This moves the optimum load resistance marked A in Figure 5.16(b) to the point B. Similarly, in moving from reference plane B to C, and looking towards the load, the series reactance of the 0.55-nH inductor needs to be removed to move from B to C, since it is part of the load in B but not in C. This move is also shown on Figure 5.16(b).

As a result of absorbing the parasitics into the measurement, the optimum load resistance has been transformed from a real resistance at reference plane A into a complex impedance at reference plane C. A similar transformation shifts the other load pull contours as well, as shown in Figure 5.16(c). These now represent the possible impedances that should terminate the device at its external terminals to produce the nominated level of output power. For comparison, the load pull contours of this device under similar conditions are simulated from the nonlinear device model in Section 5.4.2.

Of course, in synthesizing a design, it is easier to match a 50-Ω load to a real impedance 4-Ω (A), rather than to a complex impedance (C). For this reason, the matching network should attempt to match to (A) but incorporate the shunt capacitor and series inductor as the first elements of the matching network. The presence of more complicating embedding impedances can be handled by linear CAD tools, but the principle of removing the elements in moving from an intrinsic reference plane to an extrinsic one is still valid.

Real power devices, such as we have used here, have very low impedances close to the short-circuit side of the Smith chart. If we choose a smaller device, its optimum load resistance will be larger and the contours will begin to look more "circular," as shown in Figure 5.17 for $R_{OL} = 20\Omega$. If the load is changed to either half of the optimum resistance or

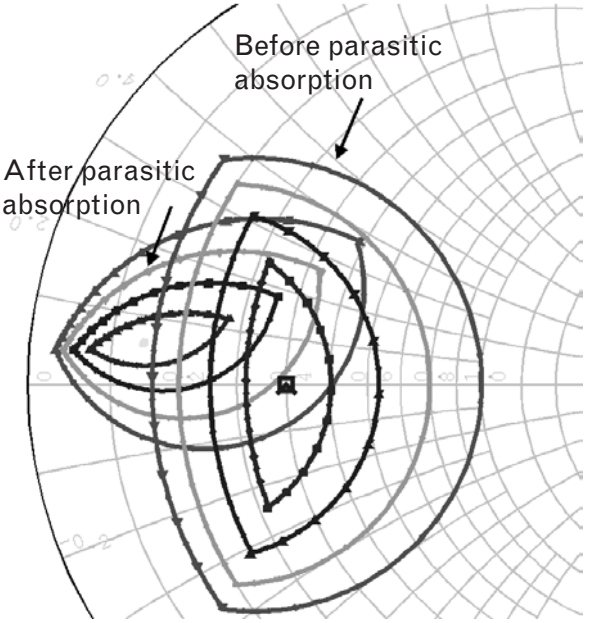


FIGURE 5.17
Approximate load pull contours (1-dB steps) for a device with the same parasitics but $R_{LO} = 20\Omega$.

conductance (i.e., 10Ω or 25 mS), the power delivered into these loads will be half of the maximum available. The same is true for all complex loads with the same real parts, and whose imaginary parts lie within the ranges specified by (5.14) and (5.17). This defines, for example, the –3-dB power contour shown in the figure.

Now consider the same exercise but at 900 MHz instead of 1,900 MHz. When performing the deembedding, the shunt susceptance to be removed is halved since the frequency is approximately halved (A to B′); likewise, the series reactance to be removed is also halved (B′ to C′). This shifts the optimum load point (C'), and the 1-dB power contour, by a smaller "rotation" on the Smith chart as shown in Figure 5.16(b).

We can now construct the locus of the optimum power load from 900 to 1,900 MHz. It moves in a counterclockwise direction along the locus shown on the Smith chart. Unfortunately, to construct a matching network to follow this locus, and thereby to achieve maximum power from 900 to 1,900 MHz, is impossible. The locus moves the wrong way on the Smith chart. Any distributed element or real element will always move clockwise on the Smith chart, except in the vicinity of a resonance where any movement will be narrowband. For instance, the impedance of any transmission line with frequency always rotates clockwise. To construct a locus to track that required for optimum power, negative matching elements are needed. This problem is similar to that encountered in conjugate matching across frequency—it is impossible to match to s_{22}^* across a broad bandwidth because the conjugate element of the shunt output capacitor of the device is a negative shunt capacitor.

Instead, we must compromise on the maximum power we wish to obtain from the device. We cannot obtain the peak power across a broad range of frequencies, but instead must back off the power requirement to a lower level. By choosing points properly (e.g., on the –2-dB contour of output power), it is possible to select a locus moving with frequency in the "right" direction that can be matched to ensure that power is available right across the band. This is one of the most useful properties of load pull power contours, in that they can be used in the same way as operating gain circles. Suitable load impedances can be chosen to maintain level output power as a function of frequency.

5.3 Categories of amplifiers

In the following sections, we will use the notation usually associated with bipolar transistors and refer to the collector, base, and emitter of the device. Unless specifically noted, the same discussion is appropriate for FET devices as well, with the appropriate changes to device terminals and notation.

5.3.1 Class-A amplifier

To this point we have, in fact, biased every device we have considered in the middle of its active region, and it remains in its active region at all times. Familiar to most engineers, this creates a class-A amplifier, which can amplify power linearly with minimum distortion. Although the topology and bias are often the same as for a small-signal amplifier, the key difference that we have seen for a class-A power amplifier is that the output network is optimized to permit a simultaneous large voltage and current swing at the output of the device.

In a class-A amplifier, by definition, the device is always biased on. We can write the total device current as

$$i_D = I_Q + I_{PEAK} \cos \omega t \tag{5.18}$$

where the total current swing $I_Q + I_{PEAK} \leq I_{MAX}$. From our load line analysis, we know that the largest value that I_{PEAK} can assume is just I_Q

$$I_{PEAK} \leq I_Q \tag{5.19}$$

since the device conduction current can never become negative. So if the device is biased midway on the I-V curves between zero and I_{MAX}, then choosing $I_Q = I_{MAX}/2$ implies that the zero-to-peak RF current swing can have amplitude up to $I_{MAX}/2$. For a load resistor R_L presented at the intrinsic terminals of the device, the device voltage is

$$v_D = V_{CC} + I_{PEAK} R_L \cos \omega t \tag{5.20}$$

We have seen earlier that if the RF output voltage swings down to the knee of the I-V curve where $v_D = V_{SAT}$, then the zero-to-peak sinusoidal amplitude of the collector voltage V_{PEAK} is just

$$V_{PEAK} = I_{PEAK} R_L \leq V_{CC} - V_{SAT} \tag{5.21}$$

The dc power is constant at $P_{dc} = I_Q V_{CC}$, the RF power is

$$P_{RF} = \frac{I_{PEAK}{}^2 R_L}{2} \tag{5.22}$$

and the efficiency is therefore

$$\eta = \frac{I_{PEAK}{}^2 R_L}{2 I_Q V_{CC}} \tag{5.23}$$

This innocuous and familiar-looking expression states that as the drive increases, the efficiency increases as the square of the output current or output voltage swing. The maximum efficiency occurs when the current and voltage swings take on their maximum values given above. Then,

$$\begin{aligned} P_{OPT} &= \left. \frac{I_{PEAK} V_{PEAK}}{2} \right|_{MAX} = \frac{I_Q \left(V_{CC} - V_{SAT} \right)}{2} \\ \eta_{MAX} &= \frac{P_{OPT}}{P_{dc}} = \frac{\left(V_{CC} - V_{SAT} \right)}{2 V_{CC}} \end{aligned} \tag{5.24}$$

Therefore, the maximum efficiency from a class-A amplifier cannot exceed 50%, and this only when $V_{SAT} = 0$. When the saturation voltage is a significant percentage of the bias voltage, the loss in efficiency can be dramatic.

Equation (5.23) also states that if the drive disappears altogether, the efficiency drops to zero, and all the dc power must be dissipated by the device. It is sometimes helpful to think in terms of the power dissipated by the device. In this case,

$$P_{DISS} = P_{dc} - P_{RF} \tag{5.25}$$

These expressions are plotted in Figure 5.18, where the horizontal axis represents the RF output voltage swing V_{PEAK} on the collector. As noted above, this changes from 0 when there is no drive to $(V_{CC} - V_{SAT})$ when the device is saturated.

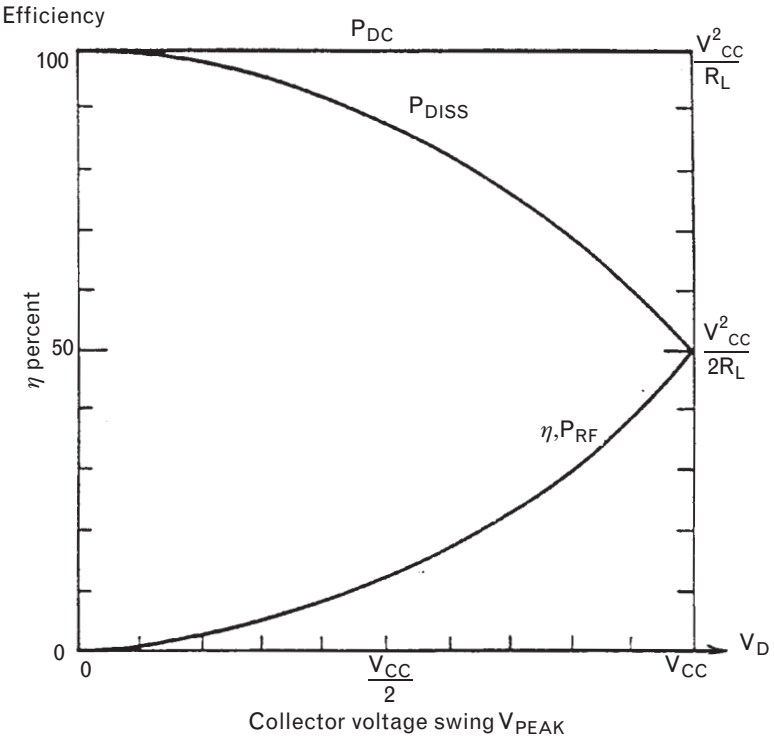


FIGURE 5.18
The efficiency, output power, dc power, and dissipated power of a class-A amplifier as a function of the RF collector voltage swing.

5.3.1.1 Example of a class-A power device

Figures 5.19 and 5.20 show extracts from the data sheet of a medium power silicon bipolar transistor that is typically driven class-A. This device, an AT-64020 from Agilent Technologies, is a bipolar transistor optimized for a high breakdown voltage (40V), and thus can attain good output power levels at modest current. If we bias the transistor with a good low-impedance path at the base, then V_{CBO} = 40V is the relevant maximum voltage swing allowed between the collector and base. This will be approximately equal to V_{CES}, which is also sometimes quoted in data sheets. Since the base voltage is typically less than 1V, the zero-to-peak collector to emitter voltage swing can be as high as 20V less the value of V_{SAT}.

If, for instance, we bias the device with V_{CC} = 16Vand a quiescent current 110 mA, and assume that the saturation voltage is 2V, then the zero-to-peak voltage swing is 14V and the output power given by

$$P_{OPT} = \left.\frac{I_{PEAK}V_{PEAK}}{2}\right|_{MAX} = \frac{0.110 \cdot 14}{2} = 0.77W$$

or 28.8 dBm. This correlates well with the saturated output power given on the data sheet in Figure 5.20, which plots more generally the relationship between output power as a function of input power, bias current, and frequency. Although the gain decreases with frequency at approximately

Features

- High output power:
  27.5 dBm Typical P1 dB at 2.0 GHz
  26.5 dBm Typical P1 dB at 4.0 GHz
- High gain at 1 dB Compression:
  10.0 dB Typical G1 dB at 2.0 GHz
  6.5 dB Typical G1 dB at 4.0 GHz
- 35% Total efficiency
- Emitter ballast resistors
- 3Hermetic, metal/beryllia package

Description

The AT-64020 is a high performance NPN silicon bipolar transistor housed in a hermetic BeO disk package for good thermal characteristics. The device is designed for use in medium power, wide band amplifier and oscillator applications operation over VHF, UHF and microwave frequencies.

Excellent device uniformity, performance and reliability are produce by the use of ion-implantation, self-alignment techniques, and gold metallization in the fabrication of these devices. The use of ion-implanted ballast resistors ensures uniform current distribution through the multiple emitter fingers.

AT-64020

200 mil BeO package

(a)

AT-64020 Absolute maximum ratings

Symbol	Parameter	Units	Absolute maximum
V_{EBO}	Emitter-base voltage	V	2
V_{CBO}	Collector-base voltage	V	40
V_{CEO}	Collector-emitter voltage	V	20
I_c	Collector current	mA	200
P_T	Power dissipation[2-3]	W	3
T_j	Junction temperature	°C	200
T_{STG}	Storage temperature	°C	−65 to 200

Thermal resistance [2-4]
θ_{jc} = 40°C/W

Notes:
1. Permanent damage may occur if any of these limits are exceeded.
2. T_{CASE} = 25°C.
3. Derate at 25 mW/°C for T_C >80°C
4. The small spot size of this technique results in a higher, though more accurate determinaton of θ_{jc} than do alternate methods. See MEASUREMENTS section "Thermal resistance" for more information.

Electrical Specifications, T_A = 25°C

Symbol	Parameters and test conditions[1]		Units	Min.	Typ.	Max.
$\|S_{21}\|^2$	Insertion power gain: V_{CE}=16V, I_C=110 mA	f = 2.0 GHz f = 4.0 GHz	dB		7.0 2.0	
P_1 dB	Power output @ 1 dB gain compression	f = 2.0 GHz f = 4.0 GHz	dBm	26.5	27.5 26.5	
G_1 dB	V_{CE} = 16 V, I_C = 110 mA 1 dB Compressed gain: V_{CE} = 16 V, I_C = 110 mA	f = 2.0 GHz f = 4.0 GHz	dB	8.5	10.0 6.5	
η_T	Total efficiency at 1 dB compression: V_{CE} = 16 V, I_C = 110 mA	f = 4.0 GHz	%		35.0	
h_{FE}	Forward current transfer ratio: V_{CE}=8V, I_C=110 mA		–	20	50	200
I_{CBO}	Collector cutoff current: V_{CB} = 16 V		μA			100
I_{EBO}	Emitter cutoff current: V_{EB} = 1 V		μA			5.0

(b)

FIGURE 5.19 *Data sheet for the AT-64020 bipolar transistor. (Courtesy of Agilent Technologies.)*

the classic rate of 6 dB/octave, the available output power falls off much more slowly, since the maximum available voltage and current swings stay (almost) the same, assuming there is sufficient input drive available.

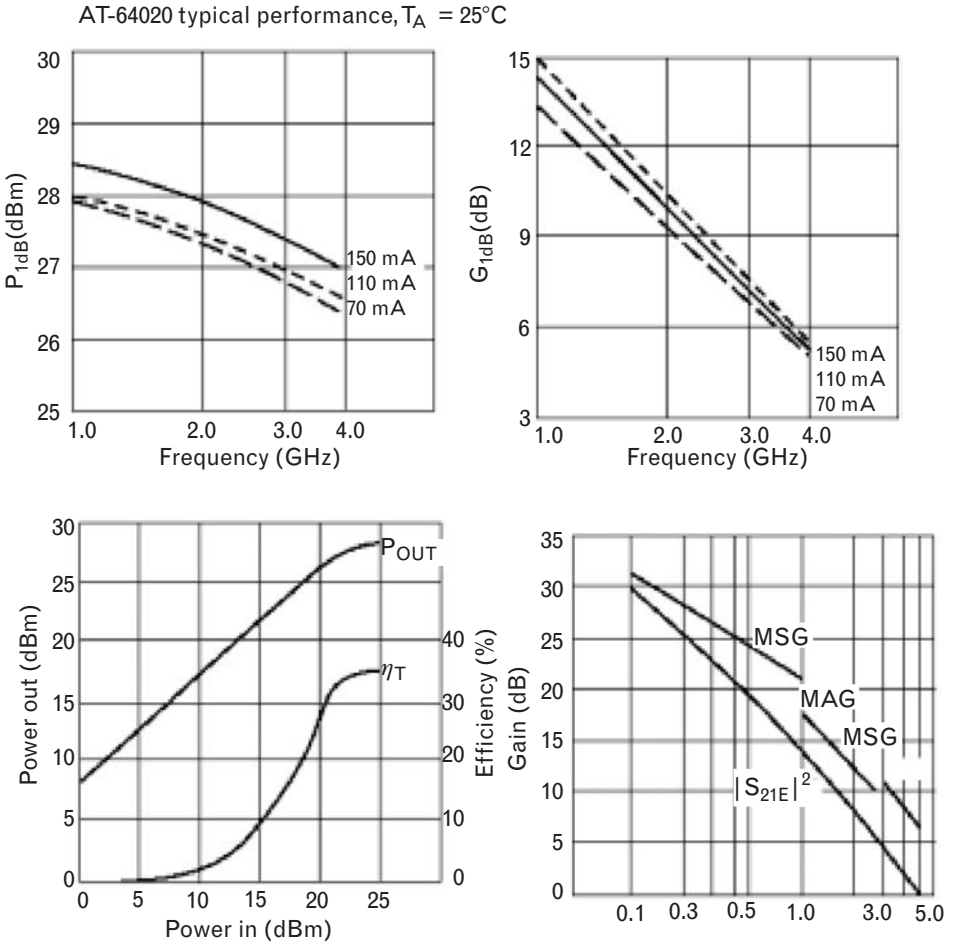


FIGURE 5.20 *Characterized data for the AT-64020 bipolar transistor. (Courtesy of Agilent Technologies.)*

Figure 5.19 states that the maximum power dissipation cannot exceed 3W at room temperature. In addition, the maximum current should never exceed the value determined by the manufacturers at which the device will be damaged by excessive current (i.e., 200-mA quiescent current for this transistor). Similarly, the dc bias voltage should never exceed one-half of the breakdown voltage V_{CBO}, so that when the collector voltage swings positive, the total voltage from (5.20) is less than V_{CBO} (i.e., 20V in this case). The safe operating area curve can be determined by these absolute limits, and by calculating those values of average collector current and voltage that keep the temperature less than the maximum safe operating temperature, 200°C. If the ambient temperature is 25°C, the power dissipated must not cause a temperature increase of any more than 175°C. Since the thermal resistance of the transistor itself is 40°C per watt, a dissipation greater than 175/40W or 4.4W would cause the average temperature to

rise above this. This is higher than the rated 3W, but there will also be hotter spots within the device, and any increase in thermal impedance caused, for instance, by poor heat sinking, would lower the allowable dissipation we have just calculated.

5.3.2 Class-B amplifier

The conduction angle of a device is the angle, measured in degrees or radians over one period, for which the device remains conducting. Thus, the conduction angle of a class-A amplifier is 360° or 2π radians. The class-B amplifier is a very special case of an amplifier because its conduction angle remains 180°, independent of drive level. To define the conduction angle mathematically in terms of applied voltages, consider Figure 5.21, which shows the transfer characteristic of a device with transconductance G. When the applied voltage V_{IN} exceeds a threshold value of V_0, the output current is given by $I_O = G(V_{IN}-V_O)$, and when the input voltage falls below the threshold, the current is zero. If now we apply a sinusoidal input voltage of peak value V_1 biased at a bias voltage V_b, so that

$$V_{IN} = V_b + V_1 \cos \omega t$$

and if $V_x = V_0 - V_b$ defines the offset of the bias voltage below the threshold, then we have conduction whenever the peak input RF voltage sinusoid V_1 exceeds V_x, that is,

$$\phi = \cos^{-1} \frac{V_x}{V_1} \tag{5.26}$$

The conduction angle is 2ϕ and defines the number of degrees or radians for which the input voltage exceeds the threshold value and the device conducts. When the device is conducting, the output current mirrors the peaks of the input sinusoidal voltage, as in Figure 5.21. The peak of the sine wave tips of the output current is just the transconductance times the amount by which the input voltage exceeds the threshold; that is,

$$I_p = (V_1 - V_x)G$$

For the class-B amplifier, $V_x = 0$, so $V_b = V_0$ and the device is biased right at threshold. For a bipolar transistor, this corresponds to biasing the base at roughly 0.7V, and for an FET, to biasing the gate at pinch-off. In this case the output current is simply a half-wave rectified sinusoid, which can be expanded as a Fourier series

$$i_o(t) = \frac{I_p}{\pi} + \frac{I_p}{2}\cos \omega t + \frac{2I_p}{3\pi}\cos 2\omega t - \frac{2I_p}{15\pi}\cos 4\omega t + \ldots \tag{5.27}$$

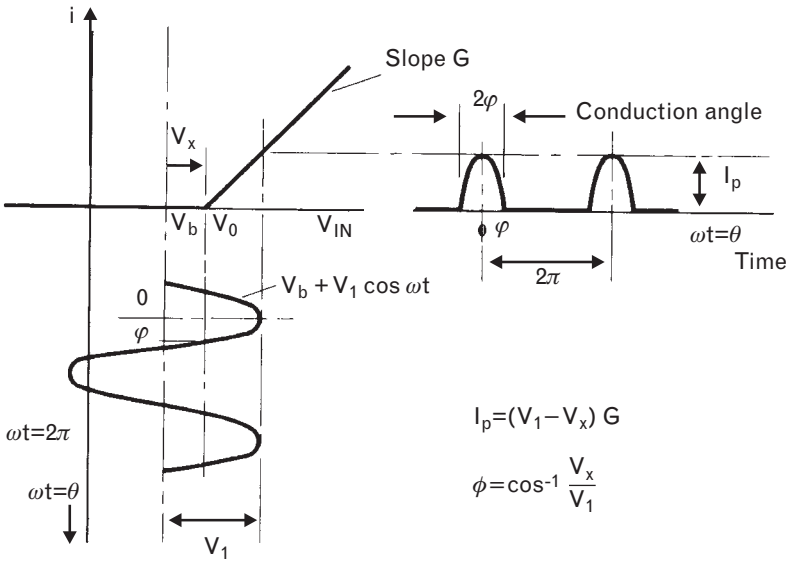


FIGURE 5.21
The transfer characteristics of a device with turn-on voltage V_o. (From: [2]. © 1971 Addison-Wesley Inc.)

If at the output of the amplifier we have a bandpass or lowpass filter to eliminate the higher harmonic components, then the output voltage will be a sinusoid of zero-to-peak value

$$V_{PEAK} = \left(\frac{V_1 G R_L}{2}\right) \tag{5.28}$$

If, instead of employing a sinusoidal waveform at the input to drive the device, we used a square wave, the output would also be a square wave and could be similarly expanded as

$$i_o(t) = \frac{I_p}{2} + \frac{2I_p}{\pi}\cos\omega t - \frac{2I_p}{3\pi}\cos 3\omega t + \frac{2I_p}{5\pi}\cos 5\omega t - \ldots \tag{5.29}$$

where I_p is the peak-to-peak amplitude of the square wave.

The key point to observe here is that in class-B operation, the conduction angle stays constant at 180° even as the input voltage is increased, and (5.27) or (5.29) are valid for all values of I_p. As a result, we can write the relationship in (5.28), and this is a linear relationship between the input voltage and the output voltage. Thus, although the device itself is nonlinear to the extent that the current waveform does (and must) contain higher harmonics, these can be removed and the relationship of the fundamental output to the input is linear since it does not depend on the input drive level. This assumes, of course, that the transconductance G is constant and the transfer characteristic linear. In real devices, G falls to zero around the turn-on point V_0. Therefore, it is common to use a bias voltage V_b slightly

above the threshold voltage V_0 so that near-linearity can be maintained, and the gain loss not as severe.

To calculate the dc power, we note from (5.27) that the dc current is I_p/π, and that the zero-peak current swing I_{PEAK} is $I_p/2$, so we have

$$\begin{aligned} P_{dc} &= \frac{V_{CC} I_p}{\pi} \\ P_{RF} &= \frac{V_{PEAK} I_{PEAK}}{2} = \frac{{I_{PEAK}}^2 R_L}{2} = \frac{I_p^2 R_L}{8} \\ \eta &= \frac{\pi I_p R_L}{8 V_{CC}} \end{aligned} \tag{5.30}$$

The output power is the same as for a class-A amplifier driven to its same limit $I_{MAX} = 2I_Q$, because then the fundamental component of the class-B collector current with $I_p = I_{MAX}$ is the same. Both have a zero-to-peak value I_{PEAK} of $I_{MAX}/2$. However, the class-B device, being more efficient, is able to run much cooler. With class-B operation, the efficiency rises linearly with the input current or voltage. If we set the load to the optimum to allow us to achieve the maximum current swing of $I_p = I_{MAX}$, and to allow the collector voltage to swing down from the supply at V_{CC} to V_{SAT} and back up to $2V_{CC} - V_{SAT}$, then the optimum load resistor is the same as given by (5.6) and the efficiency becomes from substitution into the above

$$\eta = \frac{I_p}{I_{MAX}} \left(\frac{V_{CC} - V_{SAT}}{V_{CC}} \right) \frac{\pi}{4} \tag{5.31}$$

This has a maximum value of close to $\pi/4$, or 78%, when the peak value of the output current half-sinusoid I_p achieves its maximum value I_{MAX}. This is the key advantage of the class-B amplifier over the class-A amplifier, since the increased efficiency allows considerable improvement in radio talk time. The reason for the improvement is that now the collector current is zero for half a cycle, when the output voltage is the highest. Unfortunately, we will see shortly that the gain of the class-B amplifier is less than for class-A, so the power-added efficiency is barely improved, if at all.

These expressions are plotted in Figure 5.22. As before, the *x*-axis is the zero-to-peak signal swing at the output V_{PEAK}, related by (5.28) to the input voltage swing. As V_{PEAK} increases from zero to its maximum value of $V_{CC} - V_{SAT}$, the output power increases as the square of the input while the dc power increases linearly, from (5.30). As a result, the power dissipated in the device, given by (5.25), passes through a maximum at an input power level just prior to compression.

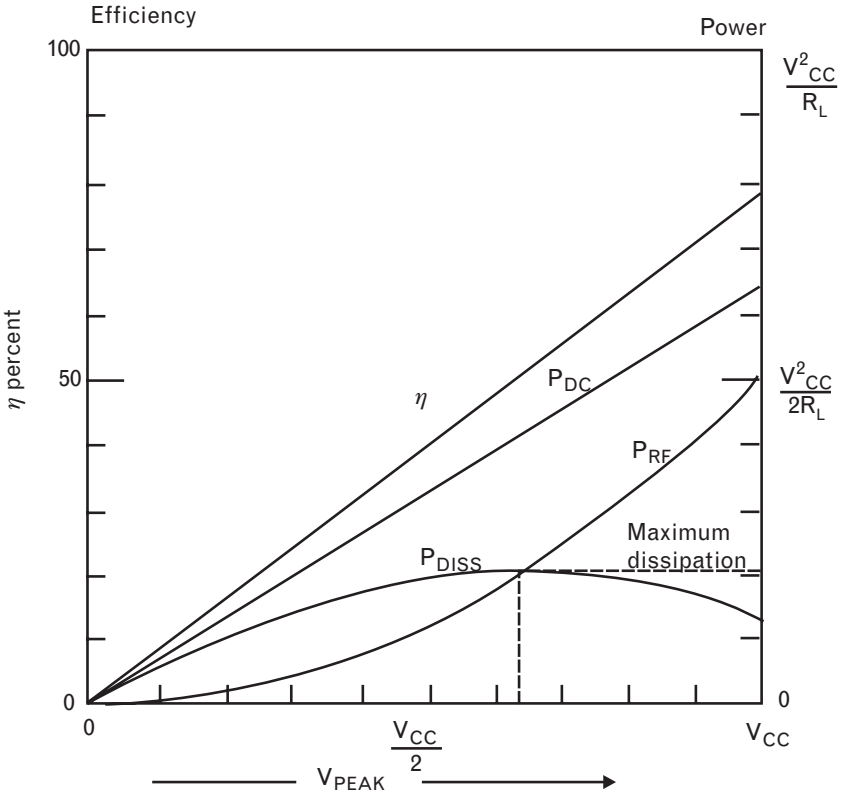


FIGURE 5.22
The efficiency, output power, dc power, and dissipated power as a function of drive for a class-B amplifier.

Figure 5.23 shows the load line for class-B operation of a FET. The bias point has shifted to B in the figure, since the quiescent current is set close to zero. The drain voltage can be kept the same as for class-A operation, although if the bias point is selected so that the drain voltage is allowed to swing up to the breakdown voltage at high input levels, it must be reduced by $V_p/2$ compared with class-A operation. This is because the breakdown occurs between the drain and the gate; as the gate bias voltage is reduced from $-V_p/2$ in class-A to $-V_p$ for class-B, this pulls the drain-gate diode closer to breakdown. The drain voltage is therefore reduced by the same amount to keep the differential voltage across the diode the same.

We stated above that the optimum load resistor for class-B is given by (5.6) (i.e., the same as for class-A). This is an unexpected result to those who would rather take the slope of the "on" portion of the load line as the optimum load resistor, which gives a resistance one-half the true value. In fact, this argument ignores the resistor value during the "off" period where it is infinite. Conceptually it is easiest to think that as long as the drain is biased through an RF choke so the voltage at the drain can "float" around its average bias value of V_{DD}, the endpoints of the class-B waveform are the same as those for class-A. Consequently, the average slope is the same, so that the optimum load resistor is also the same. Perhaps a more mathematical line of thought is that in the frequency domain, the fundamental output

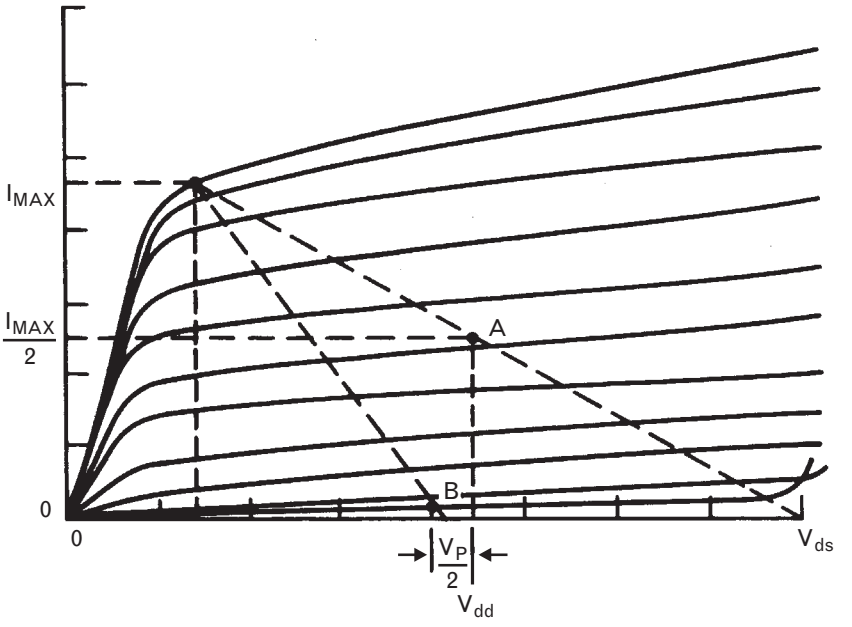


FIGURE 5.23
The load line for class-B operation in comparison with that for class-A. The curves shown are those for an FET. (From: [3]. © 1990 John Wiley & Sons, Inc. Reprinted with permission.)

current and voltage are the same for class-A and class-B operation, so the fundamental load resistance must be the same as well.

5.3.2.1 The push-pull configuration

We have seen above that a class-B amplifier is linear if the transconductance is constant and if the harmonic components of the output current are short-circuited by a tank circuit at the load, so that the output voltage contains only the fundamental component. From (5.28), this component is linear with the input voltage.

In fact, with microwave and RF power devices, the tank circuit is often omitted because the harmonics tend to be short-circuited by the device output capacitance. Although this means that we may not (necessarily) need to take as much care in considering how to terminate the harmonic components of device current as for lower frequency amplifiers, and that load pull measurements will be relatively insensitive to the harmonic load presented at the device, we will see shortly in a power amplifier example that a large device shunt output capacitance can both lower the output resistance and limit the bandwidth. However, it does mean that a tank circuit is unnecessary (and, we shall see, even undesirable) to short-circuit all the output harmonics.

Another way to remove the harmonic components of the class-B current is to force two half-sine wave pulses through the load in opposite directions, as shown in Figure 5.24. If two class-B devices are driven 180° out of phase, the current through the first device will be given by (5.27) and the second by

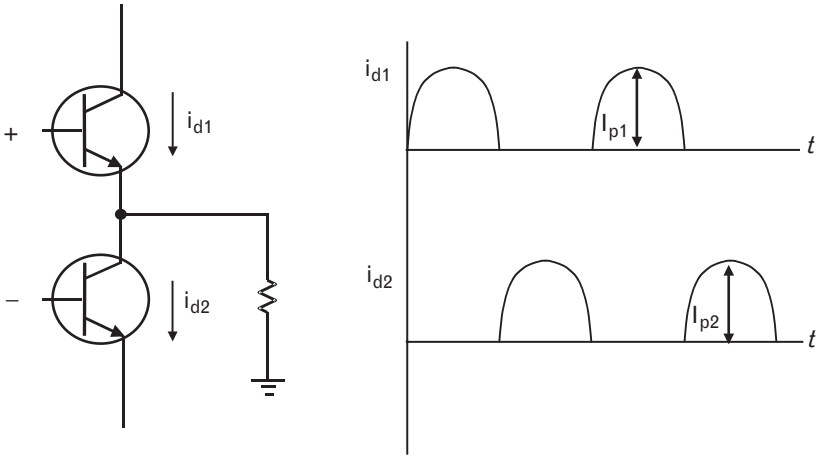


FIGURE 5.24
The principle of a push-pull amplifier is to drive two devices out of phase and subtract their output currents in the load.

$$i_o(t) = \frac{I_p}{\pi} + \frac{I_p}{2}\cos(\omega t + \pi) + \frac{2I_p}{3\pi}\cos(2\omega t + 2\pi) - \frac{2I_p}{15\pi}\cos(4\omega t + 4\pi) + \ldots \tag{5.32}$$
$$= \frac{I_p}{\pi} - \frac{I_p}{2}\cos\omega t + \frac{2I_p}{3\pi}\cos 2\omega t - \frac{2I_p}{15\pi}\cos 4\omega t + \ldots$$

If the two devices now drive the same load, but in opposite directions, their currents subtract. Subtracting (5.32) from (5.27) yields

$$i_L(t) = I_p \cos\omega t \tag{5.33}$$

which is just the output current of a class-A amplifier with twice the zero-to-peak fundamental current swing as each class-B device. Furthermore, the peak power dissipation per device is much less than for the single ended class-A stage that achieves the same output power to the load, since both devices are operated in class-B mode, and the power dissipation is shared between the two transistors.

The out-of-phase drive at the inputs is achieved with a balun that converts a single-ended input to a differential output, covered in Volume I, Chapter 7. Conceptually, it is simplest to think of this as a 1:1 transformer with its secondary terminals connected to each base. In practice it might be realized with a quarter-wave coaxial transmission line suspended above a ground plane with just the single-ended input end of its coaxial shield grounded and the remote center-conductor and shield driving each base, a special case of a 1:1 transmission line transformer. As shown in Figure 5.25, the subtraction of the currents through the same load at the output is achieved with a similar balun that converts the differential output from the two collectors into a single-ended output connected to the load resistor.

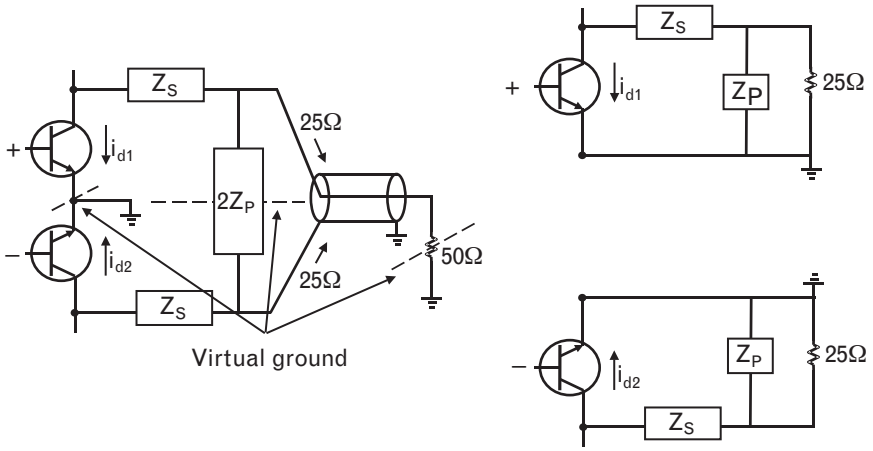


FIGURE 5.25
A symmetrical circuit with differential inputs and single-ended output using a coaxial balun. The virtual grounds are shown, allowing its separation into two identical half circuits with a real ground for analysis. Each half circuit carries half the current of an equivalent single-ended device.

Other variants of transmission line transformers with 1:4 or 1:9 transformation ratios could also be used [4]. At microwave frequencies, the rat-race coupler described later in Chapter 7, could be used. In all implementations however, the bandwidth of the balun needs to be considered, since the differential conversion may not extend to higher harmonics of the signal. If this is the case, the subtraction of two device currents of the form given by (5.32), for instance, will not cancel the harmonic components of the current and may in fact even reinforce them.

The "midpoint" of the circuit, such as the transformer center-tap, is known as a virtual ground, since no signal appears there. We first introduce this concept in Volume I, Chapter 5, when we look at impedance matching of balanced circuits. Cancellation of even harmonics is typical of such a circuit that possesses symmetry and a virtual ground. This is perhaps more of an advantage at lower frequencies, since at RF and microwave frequencies, the device output impedance is not constant over a broadband range of frequencies and the harmonics tend to be terminated by either the device parasitics and/or the matching network. Furthermore, the bandwidth of the baluns themselves may not extend much beyond the upper edge of the amplifier passband and may not even respond to higher harmonics at the output. Thus, this low-frequency theory becomes less easy to apply as the frequency increases. Even so, the virtual ground still exists due to the symmetry, as illustrated in Figure 5.25. Common-lead effects, such as emitter or source inductance to ground, that tend to reduce gain can be eliminated if the signal phases can be maintained in spite of the distributed nature of microwave circuits. The virtual ground can even be used to "ground" any shunt tuning capacitors between the two half-circuits, such as shown for Z_p in the figure, which would be realized as a single shunt capacitor of twice the half-circuit impedance.

The push-pull configuration is more commonly seen below several hundred megahertz where its implementation and benefits are more appreciated. However, it is occasionally seen at microwave frequencies because

of one other major advantage: impedance transformation. Seen at the balun terminals, the input and output impedances of the push-pull amplifier consist of the two device impedances in series. For instance, if each class-B device needs to see an output resistance of R_L, then the output balun should be loaded with $2R_L$. Furthermore, the impedance level of a single class-A device delivering the same output power would be just $R_L/2$, since it would need twice the output current of each class-B device. Thus, for the same output power, the push-pull output impedance will be four times higher than for a single device. In a 50-Ω system, this reduction to a matching requirement of 25Ω at the output of each device can be of great assistance for power devices, which typically have very low impedances. However, the baluns do not eliminate any power that may be reflected by the device and thus do not improve the VSWR. Also, if the isolation between the two parts of the balun is not good, then the circuit can be unstable as well. This is a disadvantage compared to the balanced amplifier approach, which provides both good isolation and stability.

If the currents in the two devices are not perfectly matched to each other, the subtraction performed to yield (5.33) is imperfect and residual harmonic distortion results. If the peak current I_p in the first device is I_{p1} and in the second device it is I_{p2}, the residual second harmonic distortion is given by

$$\frac{\text{second harmonic}}{\text{fundamental}} = \frac{4}{3\pi}\frac{I_{p1} - I_{p2}}{I_{p1} + I_{p2}}$$

and the residual fourth harmonic distortion by

$$\frac{\text{fourth harmonic}}{\text{fundamental}} = \frac{4}{15\pi}\frac{I_{p1} - I_{p2}}{I_{p1} + I_{p2}}$$

The push-pull implementation with baluns can appear to be quite similar to balanced amplifiers, in which power is also split equally between two balanced (equal) amplifier stages. However, the fundamental principle is quite different. Push-pull amplifiers only operate with class-B stages and require exact subtraction of output currents, achieved with 180° coupling; balanced amplifiers operate independently of the class of the amplifier and require splitting and addition of power, usually achieved with 0° or 90° coupling.

5.3.2.2 Characterization of a class-B amplifier power device

Figure 5.26 shows measured data for a GaAs MESFET that can be used to build a class-B RF amplifier. This device, the NE6500379A from NEC, can deliver 3W at 1,950 MHz when biased with 6V at the drain. From Figure 5.26, we see that the relevant breakdown voltage for amplifier operation is BV_{GD} or 17V. This corresponds to the maximum instantaneous

FIGURE 5.26
Data sheet for the NE6500379A GaAs MESFET. (Courtesy California Eastern Labs.)

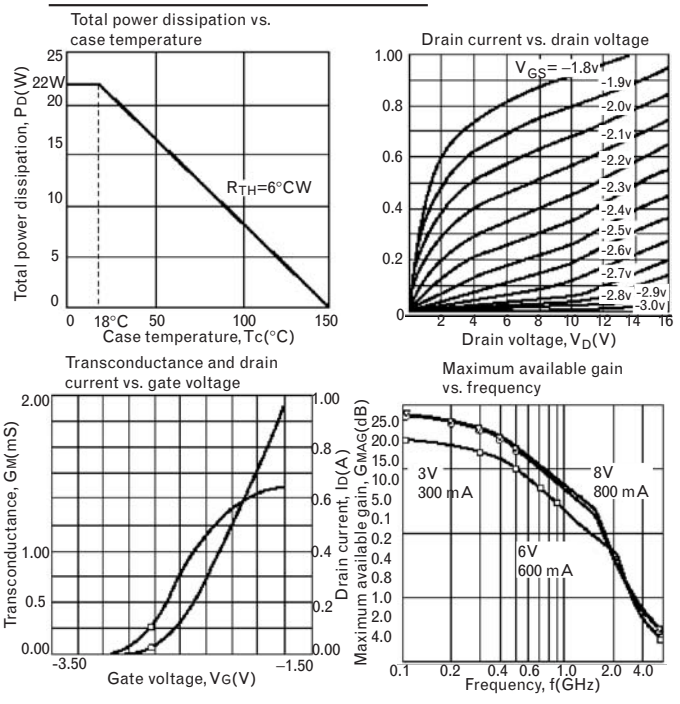
Electrical characteristics (Tc=25°W)

	Part number package outline		NE6500379A 79A				
	Symbols	Characteristics	Units	Min	Typ	Max	Test Condition
Functional characteristics	P_{1dB}	Power out at 1dB gain compression	dBm		35.0		f = 1.9 GHz, V_{DS}=6.0V Rg = 30Ω I_{DSQ} = 500 mA (RF OFF)[2]
	G_L	Linear gain[1]	dB	9.0	10.0		
	η_{ADD}	Power added efficiency	%		50		
	I_D	Drain current	A		1.0		
Electrical DC characteristics	I_{DSS}	Saturated drain current	A		4.5		VDS=2.5V; VGS=0V
	V_p	Pinch-off voltage	V	-3.6	-2.6	-1.6	VDS=2.5V; IDS=21 mA
	R_{TH}	Thermal resistance	°C/W	17	5	6	Channel to case
	BV_{GD}	Gate-to-drain breakdown voltage	V	17			I_{GD}=21 mA

Absolute maximum ratings[1] (Tc=25°C)

Symbols	Parameters	Units	Ratings
V_{DS}	Drain to source voltage	V	15
V_{GS}	Gate to source voltage	V	-7.0
I_{DS}	Drain current	A	5.6
I_{GS}	Gate current	mA	50
P_T	Total power dissipation	W	21
T_{CH}	Channel temperature	°C	150
T_{STG}	Storage temperature	°C	–65 to +150

Typical performance curves (T_A=25°C)



differential voltage allowed between the gate (negative volt) and drain (positive volt) before the gate-drain junction enters avalanche. The maximum allowed instantaneous drain current corresponding to I_p is 5.6A. This is slightly greater than the measured I_{DSS} of 4.5A and allows for the gate voltage to instantaneously swing slightly positive. Some data sheets specify the maximum *continuous* or *average* drain or collector current instead, which in class-B mode corresponds to I_p/π in (5.27) rather than I_p itself. It is important to distinguish the difference, as many data sheets do not make it clear.

Figure 5.27 shows the simulated circuit and related performance curves. The I-V curves of the device show that when the gate is biased at –2.75V (the pinch-off voltage), there is no quiescent bias current. Thus we select this gate voltage for class-B operation, and bias the drain at 6V. When the input and output matching circuits are tuned for maximum power, the fundamental, second harmonic, and third harmonic output powers achieved are as shown in the figure. The saturated output power is over 36 dBm, and the 1-dB compressed power is 34.8 dBm (3W) at an input power of 23 dBm. The second and third harmonic output powers are quite high, and in this class-B amplifier rise proportionally with the fundamental output power.

The conversion gain illustrated in the figure indicates one of the problems with the class-B approach: At low input power levels, the gain is close to zero, because the transconductance is nearly zero when the device is turned off. Only when the device is switched on for half a cycle by increasing input power does the gain increase. As the input power continues to increase, the device saturates and the gain starts to decrease again.

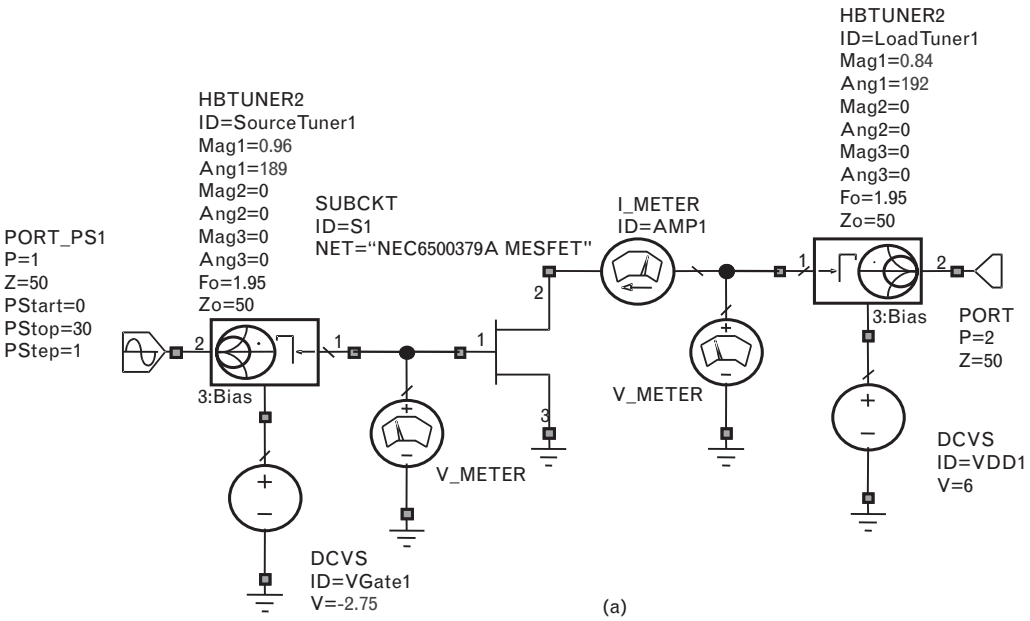
Raising the input power increases I_p, so the average bias current also increases, from nearly zero to 940 mA at the 1-dB compressed point. This is reflected in the load line, which looks more like a class-A load line at high drive levels. Since the dc component from (5.27) is I_p/π, it follows that I_p, the maximum current swing, is 2.95A at 1-dB compression. The actual load resistance seen by the device is therefore approximately 2*6V/2.95A = 4Ω, from (5.6). The efficiency at the 1-dB compression point may also be calculated as

$$\eta = \frac{P_{RF}}{P_{dc}} = \frac{3}{6 \times 0.94} = 53\% \tag{5.34}$$

consistent with the value shown in Figure 5.27. The total efficiency continues to increase with input drive up to nearly 70%, while the power-added efficiency starts to decline as the gain begins to fall.

5.3.3 Class-F amplifier

The class-F amplifier is considered as the next class of amplifier, because it is a special variant of class-B (there is no logic in the ordering of amplifier



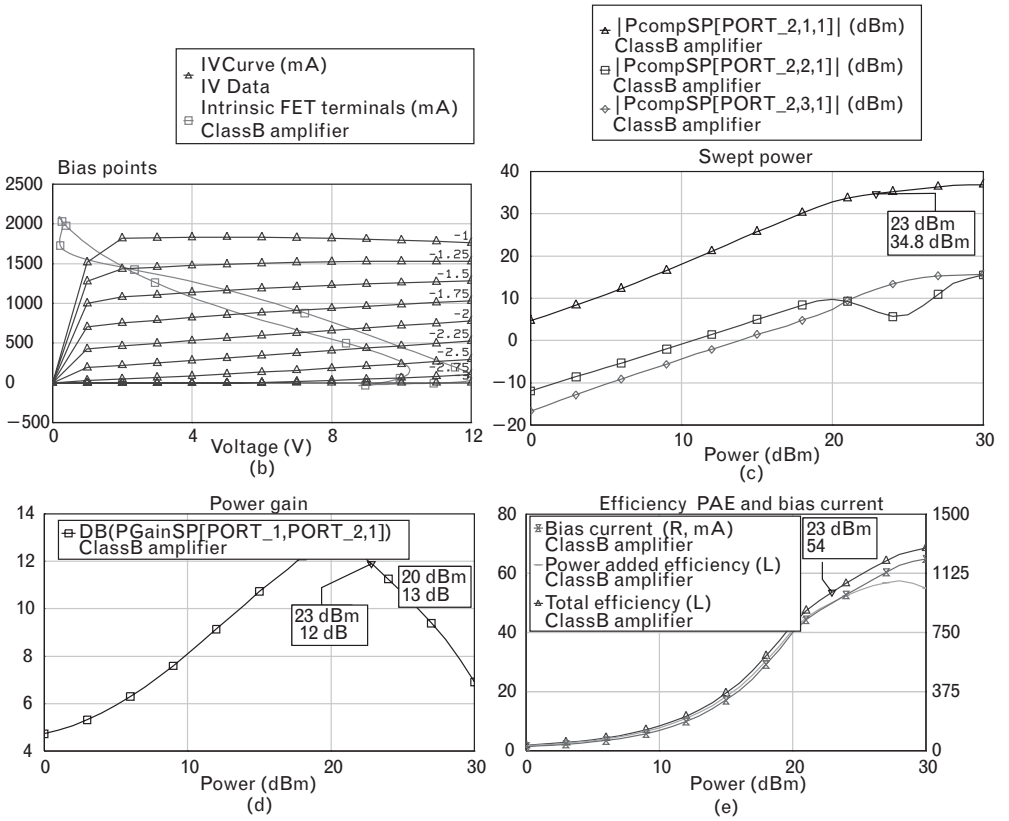


FIGURE 5.27 *(a) The NE6500379A MESFET biased and tuned by load-pull tuners as a class-B amplifier. Simulated performance curves showing (b) the load line at 1-dB compression, (c) the swept output power and harmonics, (d) the conversion gain, and (e) the drain bias current, drain efficiency, and PAE as a function of input power.*

classes). Originally considered as an option for low-frequency amplifiers, it has become more popular at RF frequencies in recent years, although the need to account for the device and package parasitics as the frequency moves higher will continue to require special effort on the part of the device designer. The class-F amplifier achieves even higher efficiency than the class-B amplifier because the output harmonics are reactively tuned, not just to ensure there is no dissipation at harmonic frequencies, but also to minimize power dissipation within the device by keeping the device voltage as low as possible when the current is high.

The transistor is driven at its input class-B. Let us first assume, as before, that the output of the transistor can be modeled by a current source with high output impedance. In this case, the output voltage can be shaped to any value independent of the current flowing (i.e., the output I-V curves are flat). By placing a third-harmonic resonator in series with the load, a third-harmonic component of voltage can be subtracted from the collector voltage waveform without affecting the class-B operation of the transistor. The effect of such a third-harmonic component in the collector or drain voltage is that the total voltage becomes flatter on its peaks, with the result that both efficiency and output power can be enhanced. As the amplitude of the third-harmonic component approaches and then exceeds one-ninth of the amplitude of the fundamental component, the resulting ripple in the total voltage becomes maximally flat, before starting to overshoot and form small troughs at the highest and lowest points in the waveform. At this point, the input power can be increased to swing the output waveform between its former limits, thereby increasing the fundamental output power as well.

Figure 5.28 shows the general principles. The input bias is still the turn-on voltage of the device. This generates the characteristic half-rectified sine wave current pulses at the output. The output voltage at the load is sinusoidal because of the presence of a tank circuit that short-circuits all but the fundamental frequency, and the collector voltage is the sum of the load voltage plus the voltage across the L3-C3 tank circuit. Since this circuit is tuned to the third harmonic, it generates a third-harmonic component of voltage so the collector voltage contains a dc component, fundamental, and third harmonic

$$v_C = V_{CC} + V_1 \sin\theta + V_3 \sin 3\theta \tag{5.35}$$

From (5.29) this is similar to a square wave. The efficiency can reach as high as 9/8 the maximum class-B efficiency, or 88%, if it can be arranged that

$$V_1 = \frac{9}{8}V_{CC}, \quad V_3 = \frac{V_1}{9} \tag{5.36}$$

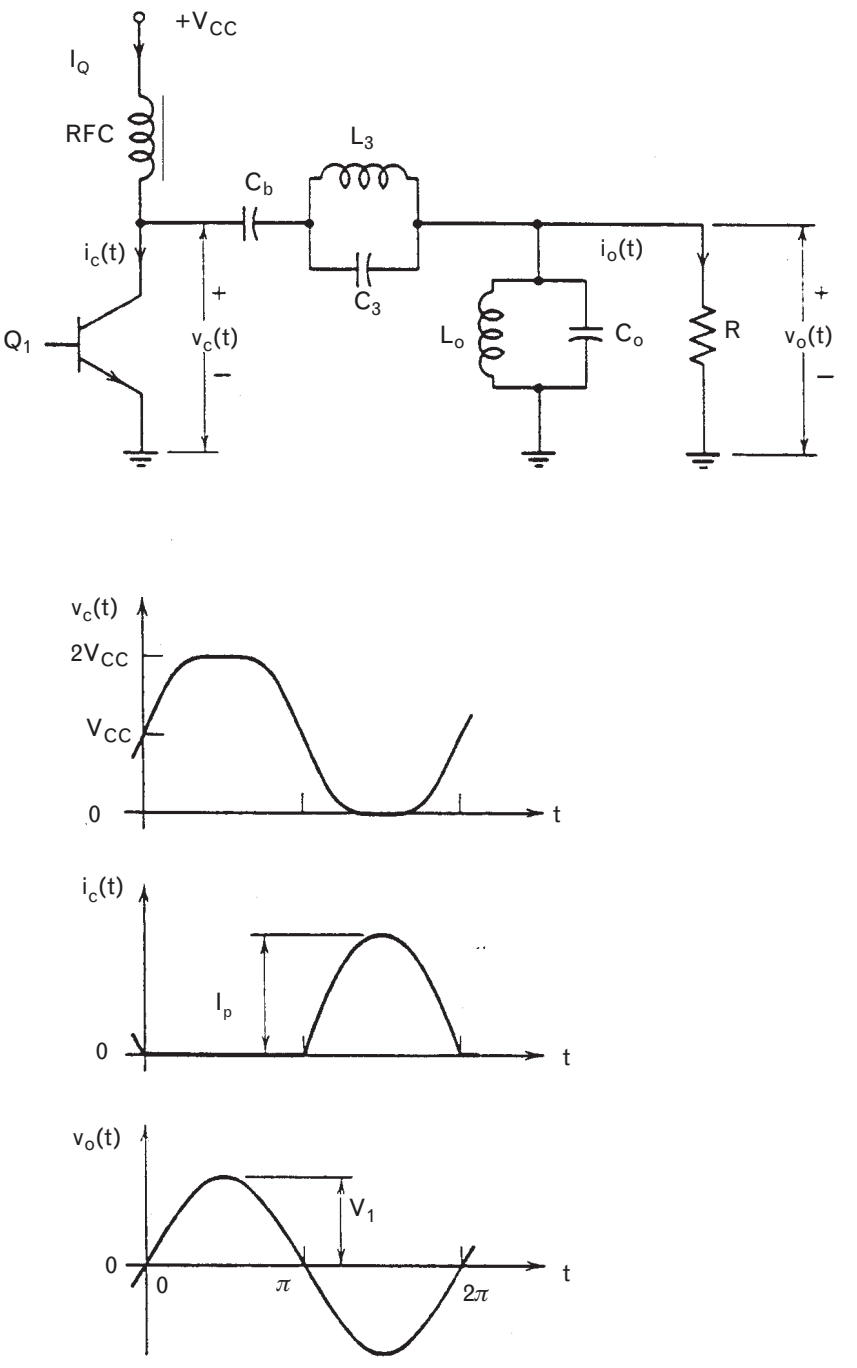


FIGURE 5.28
Principle of a class-F amplifier, showing third-harmonic tuning in the output.

In practice, this is rarely done because the principle is to shape the collector voltage to a square wave using whatever means possible. This buys a good increase in efficiency, because it can minimize the power dissipated in the device. The power dissipated in the device over one cycle is

$$P_{DISS} = \frac{1}{2\pi}\int_{\theta=0}^{\theta=2\pi} \Re e\left(i_C v_C\right) d\theta$$

The product inside the integral is already zero for half a cycle when the device is off, and can be minimized by forcing the collector voltage low when the current is high during the other half cycle. Adding the third-harmonic component to the sinusoidal load voltage lowers and flattens the voltage to reduce this contribution to the integral. The minimum value of collector voltage that can be achieved is the knee voltage V_{SAT}, and this becomes the ultimate limit to the maximum efficiency that can be achieved.

In principle, the process can be continued by harmonic tuning all higher harmonic components as well, to build up a square-wave collector voltage waveform so that the collector voltage is maintained constant at V_{SAT} throughout the entire "on" portion of the cycle. At microwave frequencies, a quarter-wave length line in series with the load circuit can be used, as shown in Figure 5.29. The transmission line can be used to transform the external load to the optimum collector load impedance for power, at the fundamental frequency. At even harmonics, the tank circuit in the load short-circuits the output, and seen through an even multiple of one quarter-wave length, presents short-circuit terminations to the collector terminals. At odd harmonics, the tank circuit in the load short-circuits the output, and seen through an odd multiple of one quarter-wave length, transforms to open-circuit terminations at the collector terminals. Thus, the collector voltage waveform is forced to consist of the fundamental and odd harmonics only, and becomes square. As a result, the only contribution to the power dissipated in the device is during the "on" portion of the cycle, and is the product of the collector current multiplied by the (hopefully low) value of V_{SAT}.

The reduced device dissipation and higher efficiency translate to higher output power at the fundamental frequency. In the case of a square-wave collector voltage with peak-peak voltage swing of $2(V_{CC} - V_{SAT})$, the fundamental frequency component by analogy with (5.29) is

$$V_{PEAK} = \frac{4}{\pi}\left(V_{CC} - V_{SAT}\right) = 1.27\left(V_{CC} - V_{SAT}\right) \tag{5.37}$$

This is 27% higher than the voltage limit when biased class-A, where the zero-peak output voltage swing is just $(V_{CC} - V_{SAT})$. Because the fundamental component of the output current remains the same, with a zero-to-peak value of $I_p/2$, the output power from the class-F amplifier is 27%, or about 1 dB higher than for class-A. Furthermore, because the device is more efficient, it can run at a cooler temperature and more reliably as well.

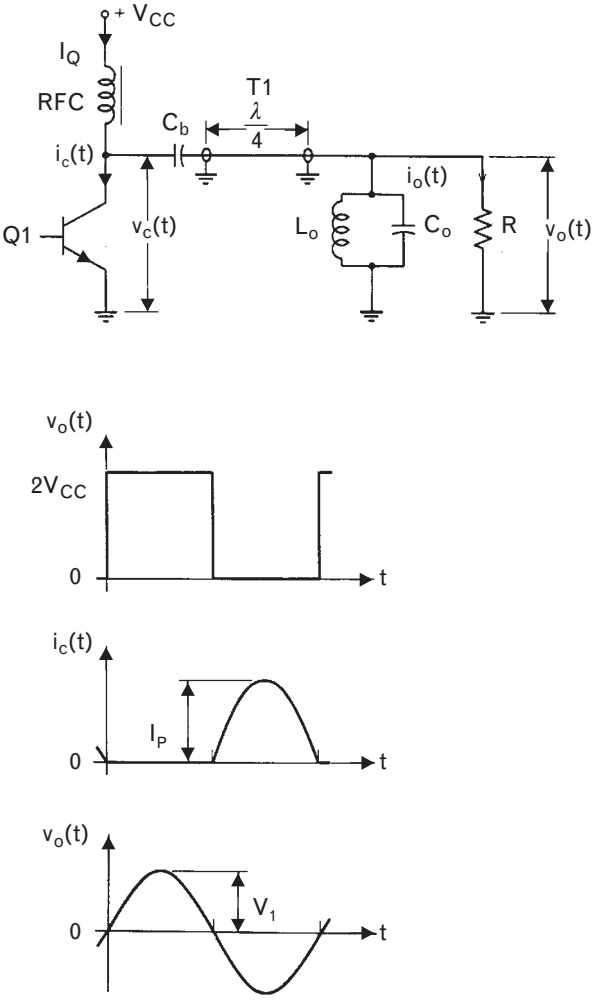


FIGURE 5.29
Transmission line "peaking" of a power amplifier to square the collector voltage waveform.

However, there are several limitations with the theory. The first is that in theory the collector current, from (5.27), contains even harmonics only. For the third-harmonic resonator to generate a third-harmonic component of voltage, it requires third-harmonic current—which does not exist. However, as the conduction angle is reduced below that of class-B (180°), the third-harmonic current rises quite strongly and this objection is overcome. In fact, third-harmonic current will also be present due to distortion generated within the transistor, but it will be weak. The second problem is that the transistor does not always behave as a *current* source, and in fact with its series collector resistance and its shunt output resistance can even be modeled as a voltage source. In this dual case, the current should be tuned to be a square wave through shunt harmonic load elements, while the output voltage should be driven class-B between the supply and ground in a half-wave rectified form. This type of operation, where the collector current is shaped as a square wave by a high impedance second-

harmonic and low-impedance third-harmonic termination is known as inverse class-F or current-mode class-F operation. Thus, in practice, it is necessary to examine the device model to understand its shunt and series RF resistances to select the appropriate operational mode. Although the theory of class-F operation may be limited in actual implementation, the principles are still widely applied to achieve useful efficiency increases of several percent over class-B.

5.3.3.1 Class-F amplifier example

A class-F FET amplifier was built using a Thomson HP07 device [5]. The design challenge was to build an efficient amplifier at a drain voltage of 3V. The problem faced at this level of V_{DD} is that with typical values of V_{SAT} (the minimum V_{DS} swing), of the order of 0.5V to 2V, the RF drain voltage swing $V_{DD} - V_{SAT}$ is considerably reduced, hurting both the output power and the efficiency in (5.31). This effect is made even worse if the optimum load line is chosen, because V_{SAT} increases as the current swing increases. This is shown by Y_{load1} in Figure 5.30.

By choosing a higher output resistance, the load line become less steep and V_{SAT} reduces. This can have a great improvement on the amplitude of the drain voltage swing $V_{DD} - V_{SAT}$ and the efficiency, particularly if V_{DD} is small to begin with. Unfortunately, the current swing is reduced and the output power suffers proportionally.

The solution suggested in [5] is to use a larger device, with greater gate length, so that g_m is increased and the original current restored. Thus, in Figure 5.30, with an increased load resistance Y_{load2}, the original power is

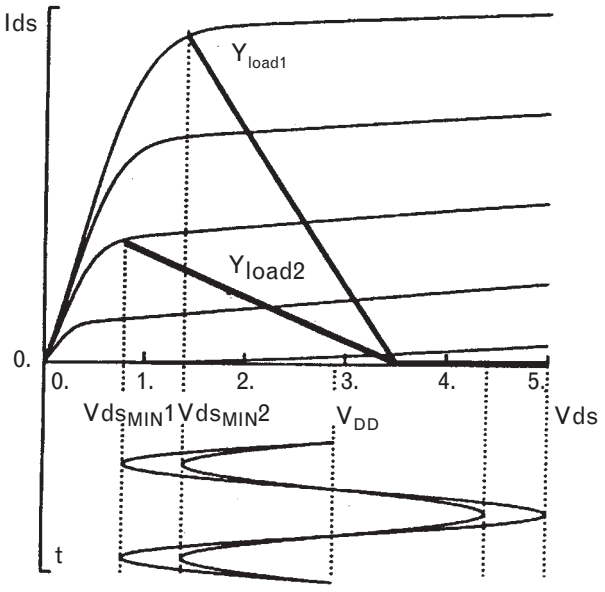


FIGURE 5.30
Reduction in V_{dsMIN} (i.e., V_{SAT}) and improvement in efficiency by reducing the slope of the load line. (From: [5]. © 1993 IEEE. Used with permission.)

restored by rescaling the current axis. Of course, the increase in the voltage swing with Y_{load2} also tends to increase the RF output power, although not sufficiently on its own to compensate for the loss in current swing.

Adjustment of the load line resistance in this way is quite a general technique, and small adjustments in R_L can result in different trade-offs between power, efficiency, and linearity [6]. Using a higher load resistor in a voltage-limited regime results in higher efficiency, but the onset of gain compression will occur earlier than for lower R_L, current-limited regimes. Depending on the dynamic range required of the amplifier, the use of a higher load-resistor can be an acceptable trade-off for efficiency against the onset of 1-dB compression.

Finally, the output-matching network is designed to match the FET output to 50Ω. Three microstrip lines are used, the first a quarter wavelength short-circuited stub at the fundamental frequency, at the drain. This has no effect at the fundamental frequency since the quarter wave transforms the short circuit to an open-circuit. The stub can also be used to bring in the drain bias, above the dc blocking capacitor that RF short-circuits the stub to ground. At the second harmonic, it presents a short circuit to the drain since the stub is effectively one half-wavelength long at that frequency. This is shown in Figure 5.31, which presents the layout of the different matching elements.

The second line is a quarter-wave open stub at the third-harmonic frequency. The resulting short-circuit at this frequency is transformed to a high impedance level at the drain terminals by the third line in series with the drain. As a result, the output voltage waveform at the drain is designed to be rich in fundamental and third harmonic, approaching a square wave in shape. The resulting waveforms are shown in Figure 5.32.

The device is biased at pinch-off (–3V) to hold the device off with no drive. The drain bias is 3V. The RF gate-source voltage at 1.8 GHz is sinusoidal about pinch-off and turns the device on for half a cycle, as seen from the half-wave nature of the drain current, which peaks at 0.36A. The zero-peak value of the fundamental component of the drain current is

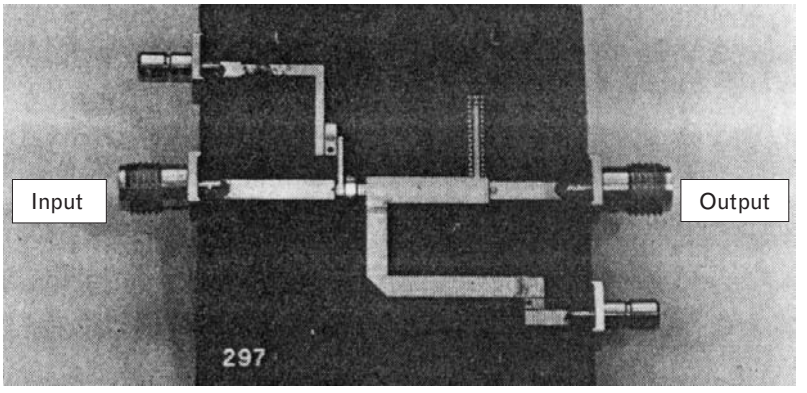


FIGURE 5.31
Layout of the class-F amplifier. (From: [5]. © 1993 IEEE. Used with permission.)

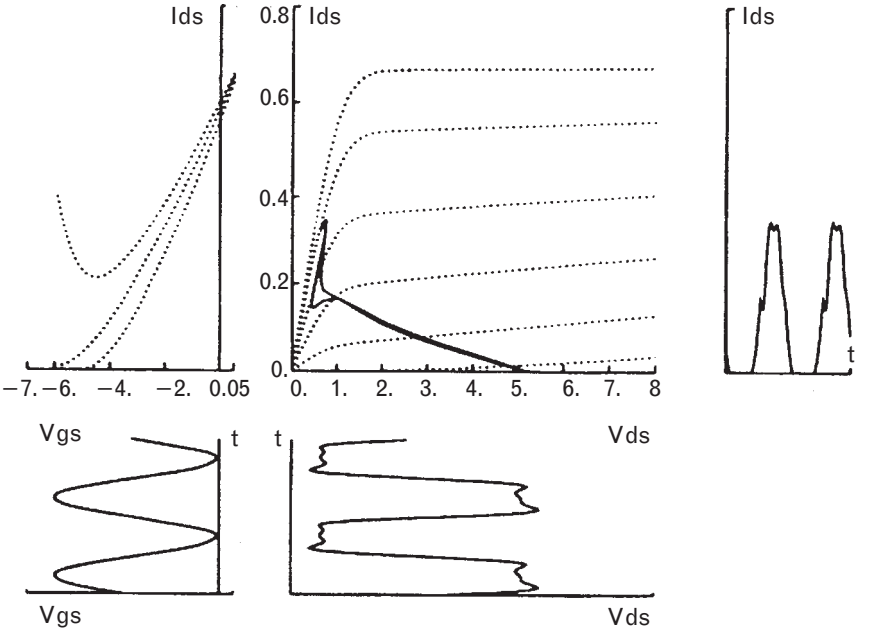


FIGURE 5.32
Simulated waveforms for the class-F amplifier showing the drain current, the load line, and the gate and drain voltages of the FET. (From: [5]. © 1993 IEEE. Used with permission.)

therefore 0.18A. The drain-source voltage swings between the saturation voltage that is 0.5V for the load line selected, and 5.5V, giving a zero-peak signal swing of 2.5V. The "dynamic" load line has shifted from the "static" load line in Figure 5.30 as a result of the dc current that is generated due to the half-wave rectified nature of the drain current. The effect of the third harmonic can be seen in the rippling square-wave nature of the drain voltage and in the contortion of the dynamic load line when it swings down to the saturation voltage. Then, the drain voltage is held constant around V_{SAT} while the drain current changes harmonically about its peak. The output power can be estimated from

$$P_{RF} = \frac{I_{PEAK} V_{PEAK}}{2} \tag{5.38}$$

as 0.18A(2.5V)/2 = 230 mW or 23.5 dBm. This is close to the measured value of 24.5 dBm. The power-added efficiency was measured to be in the 60% range, a very good value, indicating the advantage of these design steps.

5.3.4 Comparison of class-A, class-B, class-F, and other operational modes

The load lines for each of the amplifier classes we have studied are shown in Figure 5.33, drawn on the I-V curves of a FET. These show the case when the input is large enough to drive each amplifier to its maximum linear excursion along the load line, to generate maximum linear power.

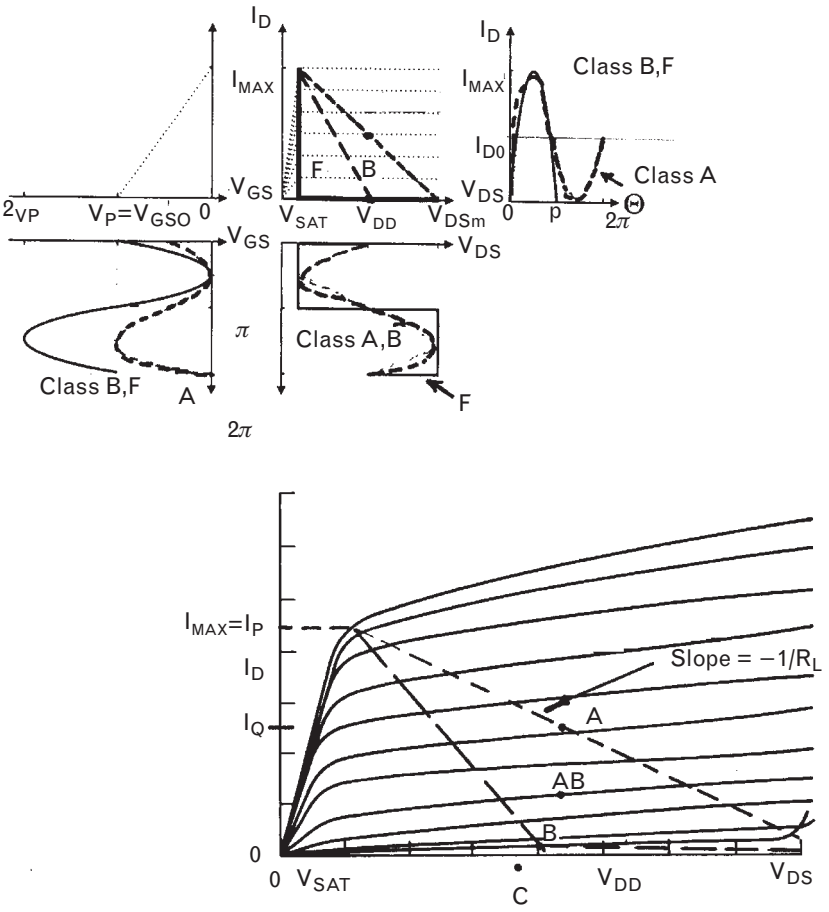


FIGURE 5.33
A comparison of the load lines for the class-A, class-B, and class-F amplifiers.

In all cases, the drain of the device is biased at V_{DD} through an RF choke, and swings between the saturation voltage V_{SAT} and its maximum value, V_{DSm} or $2V_{DD} - V_{SAT}$.

In the class-F case, the presence of a third and higher odd-order harmonic components shapes the output voltage to be square.

In class-A operation, the drain current varies sinusoidally about its quiescent value with a zero-to-peak value of $I_{MAX}/2$. In class-B and class-F operation the output current is half-sinusoidal. Because the peaks of the sinusoids are still I_{MAX}, the zero-to-peak value of their fundamental frequency component is $I_{MAX}/2$, the same as for class-A. As a result, the output power for class-A or class-B operation is the same, while as noted earlier, class-F enjoys a 1-dB higher output power, since the fundamental component of its square wave voltage output is $4/\pi$ greater than the peak value of the actual voltage swing itself.

Of course, to achieve this, the required class-B or class-F input power is four times as large as for class-A. The gate voltage swing required is doubled because the bias point is shifted to turn the device off when there is no

signal swing. Assuming equal impedance levels, the overall gain of the class-F compared with the class-A amplifier is given by

$$G_F = \frac{\frac{4}{\pi} P_{OUT,A}}{4P_{IN,A}} = \frac{1}{\pi} G_A \tag{5.39}$$

which is just under 5 dB less. The gain reduction is actually worse than this, however, because g_m is also zero around pinch-off, rather than constant as was assumed in deriving the half-sinusoidal output current from the sinusoidal input voltage. This can be observed from the reduced spacing between the FET I-V curves near pinch-off.

5.3.4.1 Class-AB amplifier

At RF and microwave frequencies where the device gain is low even when biased at class-A, it is intolerable to lose this amount of gain, so, in practice, a class-AB bias is employed, somewhere between A and B operation. This ameliorates the gain loss associated with pure class-B operation while still reducing the device dissipation associated with class-A operation, since the quiescent current is lower. A class-AB amplifier is defined by a conduction angle that lies between 180° and 360°, so the device is switched off for a portion of a cycle when the input voltage swings sufficiently into cutoff.

Cripps [6] shows that the class-AB mode can actually be more linear than even the class-A mode over a wider dynamic range of input powers, because the reduction in g_m at low bias levels (due to the square-law term in the transfer characteristic) and its saturation at forward conduction tend to compensate each other. Class-AB operation also has a slightly higher fundamental output power than class-A, assuming the same maximum current swing I_p (or I_{MAX}). Since the fundamental (zero-to-peak) component of the current swing in both class-A and class-B modes is the same $I_{PEAK} = I_p/2$, it increases slightly (by a fraction of a decibel) to pass through a maximum between the two modes as the conduction angle is reduced through class-AB operation.

These comments do assume that the input and output impedances stay the same, since the power depends not only on the voltage swing but also the impedance level. This is more true in an FET than in a bipolar device, where the increasing drive on the base-emitter junction takes it into forward bias and decreases its input resistance. As a result, the BJT suffers less gain loss than an FET when biased class-B.

Another departure from the theory is that in microwave circuits, the tank circuit at the load is rarely explicit. The function of this circuit is to short-circuit all harmonics other than the fundamental at the load. In practice, this corresponds to ensuring the harmonic powers at the output of the

amplifier are below some desired specification. Usually, the output matching circuit has such a bandpass characteristic that this is achieved through default, if not deliberate design. The output capacitance of the device C_{DS} or C_{CE} and other parasitics will also help achieve this function. However, for finite output capacitance the second harmonic component of voltage is not zero, and will be 90° out of phase with the fundamental because of the phase lag introduced by the capacitance. This makes the output (drain or collector) voltage more peaked, and with sufficient drive can cause the voltage to swing even below zero on negative half-cycles, or in any case to less than V_{SAT}. On the I-V curves of the device, this can only occur if the current simultaneously collapses towards the origin (zero) at this point. This, of course, reduces the output power and efficiency compared with the ideal (tank circuit) case.

With no tank circuit and where the output capacitance is small—as might occur if a high-f_T device is used at a very low frequency—the degradation in output power and efficiency can be several decibels. For instance, if the output capacitive shunt reactance X_C at the fundamental is around twice the load resistance (instead of the ideal zero), the best efficiency from a class-B amplifier is degraded from 78% to around 60%, with corresponding degradation in maximum output power [6]. The efficiency of a class-AB amplifier also degrades to around 60%. The degradation becomes progressively worse as the harmonic components are short-circuited less at the collector or drain as the output capacitance becomes smaller (increasing values of shunt reactance). The remedy may seem counter-intuitive, but it is to use a higher output capacitance to supplement that of the device, without perturbing the fundamental match to the load. Unfortunately, adding shunt capacitance at the output will further reduce the real part of the output resistance that needs to be matched, and reduce the available bandwidth. This capacitance should therefore be incorporated as the first element in the output-matching network to the load, so the fundamental impedance match to the ideal load resistance is still maintained.

5.3.4.2 Class-C amplifier

We have not discussed class-C amplifiers because solid-state class-C amplifiers are rare at microwave frequencies. They are a logical extension of the class-B amplifier, in which the conduction angle is reduced below 180° so that the collector or drain current is the peak of a sinusoid for a fraction of a cycle. As the conduction angle is decreased, the efficiency increases because the device is turned on for less time; however, the output power also decreases. Because the conduction angle is now a function of input drive, the fundamental component of the output current is no longer a linear function of input drive. Class-C amplifiers are therefore best used for constant envelope modulation schemes, such as FM or filtered FSK.

Some precautions are necessary when using a bipolar transistor class-C in particular. The bias voltage (hence conduction angle) can be hard to control because the transistor develops a self-bias as the large dc component of base current creates a negative voltage drop across the internal base spreading resistance. Also, a so-called *mixed-mode* termination using a series RLC output filter is sometimes preferred to a shunt RLC tank circuit. The reality of a bipolar is that its equivalent output circuit is quite nonlinear and its output resistance and large shunt output capacitance form an RC filter that prevents the collector voltage being sinusoidal. Using a series RLC filter recognizes that a sinusoidal collector voltage cannot be achieved with such an output, and instead attempts to extract the sinusoidal component of output current in the load.

5.3.4.3 Harmonically controlled amplifiers

The gain-loss inherent in class-B or class-F operation is frequently increased by shifting the input bias to the class-AB operating point. However, the gain can also be restored by arranging to achieve the same device output signal swings for a reduced input signal swing. The *harmonically controlled amplifier* (HCA) uses either a *rectangular* (rHCA) or *half-sinusoidal* (hHCA) gate voltage waveform to achieve this. Such an input voltage waveform can achieve the same output current as for class-B (or F), but without the usual large negative input voltage swing below pinch-off inherent in those amplifiers when the device is switched off.

The input and output characteristics of the hHCA amplifier are shown in Figure 5.34. The output voltage swing is the same as for class-F operation. However, the gate voltage is a half-sinusoidal waveform that drives the device from pinch-off to forward saturation during half a cycle,

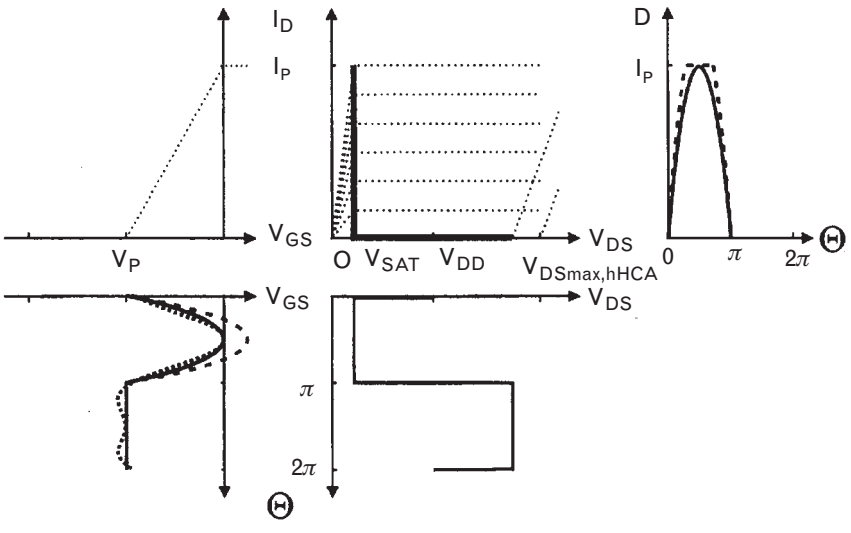


FIGURE 5.34
The dynamic load line of an hHCA amplifier, showing shaping of the gate voltage and the resulting drain voltage and current. (From: [7]. © 1999 IEEE. Used with permission.)

and keeps the device cutoff during the other half-cycle, *without* swinging the gate deep into pinch-off. Such a waveform can be approximated by adding to the fundamental frequency a second-harmonic component that subtracts out the negative half-cycle. This can be generated by combining the output voltage from a linear driver amplifier with the output from a resistively loaded class-B amplifier, using the half-sine wave current pulses at the resistive class-B output for the subtraction. Then, in the same way as for class-F, the hHCA device will create the usual half-sine wave current pulses at its output, and its voltage waveform is shaped to be square by its harmonic load impedances. However, now the input signal swing is smaller. Because the fundamental component of the input voltage is $V_p/2$ rather than V_p as for the class-F amplifier, the gain increases by up to 6 dB over class-F operation and restores the power-added efficiency by several percent. Just as importantly, because the *average* input bias point is no longer at pinch-off but shifted into the active region of the device, the intermodulation distortion can rival that of class-A amplifiers.

Harmonic control of the input voltage can also improve device reliability because the gate-drain voltage is not drawn into such a deep reverse bias (or even breakdown) during the negative half-cycle [8]. Should reverse breakdown occur, small peaks of breakdown current will flow between the drain and the gate during the negative half-cycle and drastically lower efficiency since these current peaks are in phase with the high drain voltage swing.

As already noted for mixed-mode terminations in class-C amplifiers, inverse class-F harmonic terminations—where the collector or drain current is shaped to be a square wave and the *voltage* half-sinusoidal—are also used [7]. In an rHCA, the input voltage waveform is forced to be a simple square wave swinging between pinch-off and up to forward conduction to turn the device on and off on alternate half-cycles. As before, the device achieves the very high efficiencies of class-F because one of the output voltage or current are high on alternate half-cycles and zero the remainder of the time. Using an inverse termination and shaping the collector or drain *voltage* waveform to be half-sinusoidal (rather than the current) has the added advantage that the average, or dc drain voltage can be up to one-third lower for the same output power level. This is advantageous in mobile, battery-operated conditions that require low bias voltages, or in applications where the drain voltage can be subsequently increased to obtain larger output powers.

The possibility of tuning the harmonics at both the output (to improve drain efficiency and power) *and* the input (to improve gain and power-added efficiency—and even reliability) gives a number of termination combinations that could be considered for most applications. However, for a linear input-output relationship between signal envelope, the device must be biased at pinch-off to ensure the device remains off for exactly half

a cycle, independent of input signal swing amplitude. If we are amplifying signals whose amplitude is constant, then switching-mode amplifiers present yet another alternative.

5.3.5 Switching-mode amplifiers

The ideal square-wave voltage developed at the output of the class-F transistor is identical to that across a switch, whose fundamental component of current is linearly proportional to the input voltage. The transistor has *multiple* transconductance states depending on the amplitude of the controlling (input) voltage.

The switch-mode amplifier intentionally drives a transistor as a switch between *two* binary states—either on or off. However, any linear analog relationship between the input and output voltage will then be lost. In signals without analog amplitude information, such as FM, pulsewidth modulation, or phase-varying signals with (near) constant envelope such as GMSK, the amplitude is not of concern as the information is contained in the frequency, zero crossings, or phase of the output signal and can be preserved. For signals that are amplitude sensitive, linearization techniques such as envelope restoration (see Section 5.6.6) can be used with switching-mode amplifiers, whereby the baseband amplitude information is modulated on the amplifier supply voltage and the phase variation contained within the RF signal. The output voltage of the switching-mode amplifier then scales linearly with the supply voltage and the modulation is restored. The advantage of using a switching-mode approach is that such amplifiers have drain efficiencies that can approach high values. However, careful consideration of harmonic behavior is necessary in order to avoid unintentional power dissipation at other frequencies. In the absence of any harmonic tuning, a switching-mode amplifier can dissipate up to 20% of its dc power in harmonics.

As the frequency increases, it becomes more difficult to build harmonically tuned switching-mode amplifiers and achieve worthwhile efficiencies because the output capacitance and inductance of the device introduce switching losses and make implementation of the correct tuned circuits at the intrinsic reference plane more difficult. Nonetheless, approximations to class-D or class-E amplifiers have been built in frequencies up to X-band. Raab [9] has found that the number of harmonics used to control and "build" the desired output waveform determines the maximum attainable efficiency, while it is the harmonic impedances themselves that determine the power-output capability.

5.3.5.1 Class-D amplifiers

Class-D amplifiers utilize the switching principle by driving two devices in tandem, to ensure that when one device is conducting the other device is

forced off. In effect, it can be regarded as a push-pull Class-F power amplifier. As we saw earlier, in a push-pull configuration the two devices provide paths for each other to cancel out the even harmonics. Such amplifiers have been demonstrated at 900 MHz with powers over 29 dBm and over 70% power-added efficiency [10].

The principle of the *voltage-mode* class-D amplifier is illustrated in Figure 5.35, using MOSFET devices. A differential input signal is required so that two transistors, each biased class-B, are driven 180° out of phase. A series tank circuit tuned to the fundamental is used to develop a sinusoidal *current* that switches alternately between ground and supply on each half-cycle through the corresponding on-device. If the transistors are driven sufficiently hard, the collector voltage across each device becomes square as they switch alternately between the knee voltage and the supply. However, because the transistor output capacitance must charge and discharge through a finite resistance, the voltage cannot be perfectly square. There is an output RC time constant that causes the voltage across the on-device to discharge slowly to ground or the off-device to charge slowly to the rail voltage, when they transition between off or on. A transient current spike results and some overlap of current and voltage is inevitable. When the switch closes, if V_O is the instantaneous voltage reached across the output capacitance $C_O = C_{DS}$ or C_{CE} of the off device, the energy lost is $1/2 C_O V_O^2$ each cycle. As the frequency increases, the cycle time becomes comparable

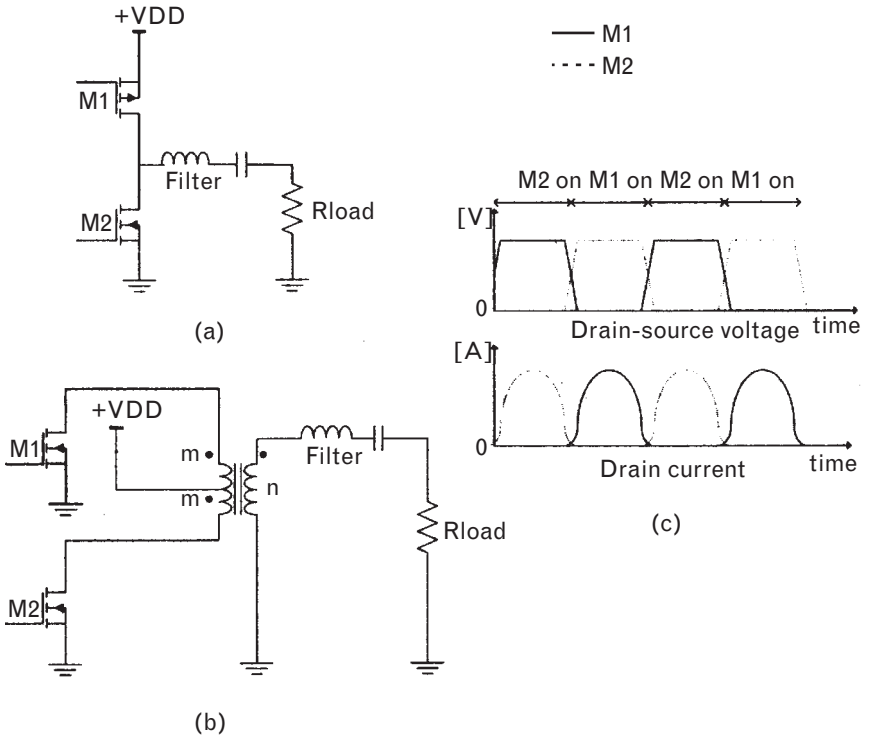


FIGURE 5.35
(a) Principle of the voltage-mode class-D amplifier operation, and (b) a more practical implementation, showing (c) drain voltage and current waveforms. (From: [10]. © 2001 IEEE. Used with permission.)

with the discharge time constant, and the switching losses become dominant.

An approach more amenable to integration because it uses a differential configuration to eliminate the need for the center-tapped transformer is to use a *current-mode* class-D amplifier, as shown in Figure 5.36. In this configuration, a constant current is shared between the two differential devices, which are switched alternately on and off. Now the voltage is kept sinusoidal by a parallel tank circuit and the current waveform kept square by virtue of the constant current source that drives the two devices. Because of the resonance of the tank circuit, the voltage at the switching instants can be forced to zero, and any output capacitance of the devices can be incorporated into the resonant load itself. Capacitive switching loss is therefore eliminated. However, the current at the switching instants may now no longer be zero, and the switching losses are given instead by $1/2Li_C^2$ where i_C is the near-zero current as the switch opens. This can be minimized by layout of the two devices to keep the series drain inductance low.

As for the class-F amplifier, harmonic terminations at the collector or drain are important to improve both efficiency and output power. Ideally, to support the square-wave like *current* waveform, the optimum load will present a high impedance at the second harmonic and a short circuit at the third harmonic. If we neglect further harmonics and force the derivatives of the current and voltage to be zero at the switching instants, then we can

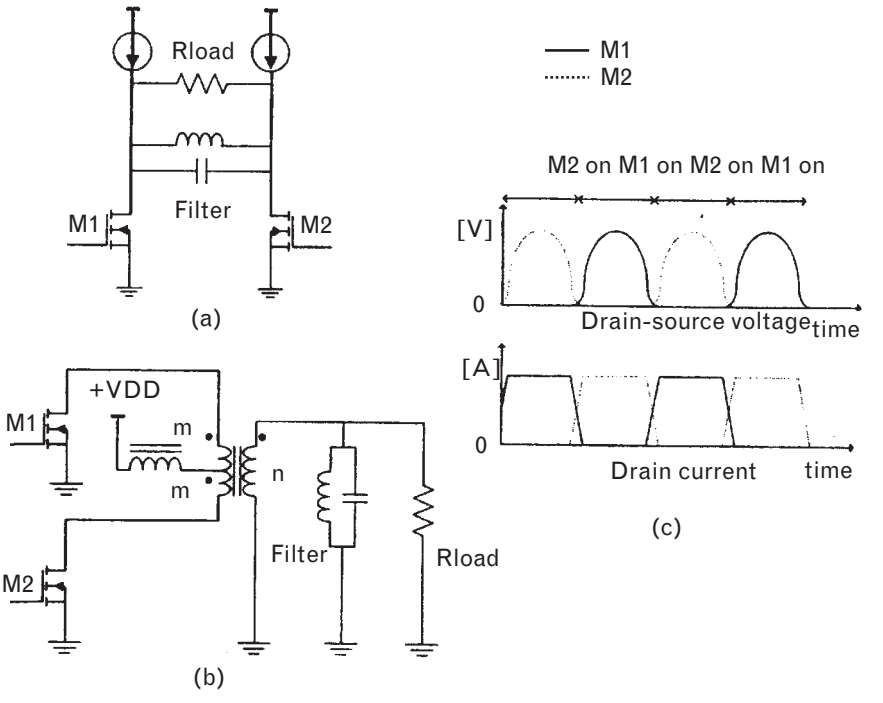


FIGURE 5.36
(a–c) Principle of the current-mode class-D amplifier to avoid switching losses by eliminating capacitive voltage at the switching instants. (From: [10]. © 2001 IEEE. Used with permission.)

prove [11] that the zero-to-peak components of the maximally flat collector voltage and current at the relevant fundamental components will be

$$V_1 = \frac{4}{3}V_{CC}; \quad V_2 = \frac{1}{3}V_{CC}$$
$$I_1 = \frac{9}{8}I_{DC}; \quad I_3 = \frac{1}{8}I_{DC} \tag{5.40}$$

The waveforms of the drain or collector voltage and current are as shown in Figure 5.37, and show the flattening of the current waveform as progressively more harmonics are included in tuning the output

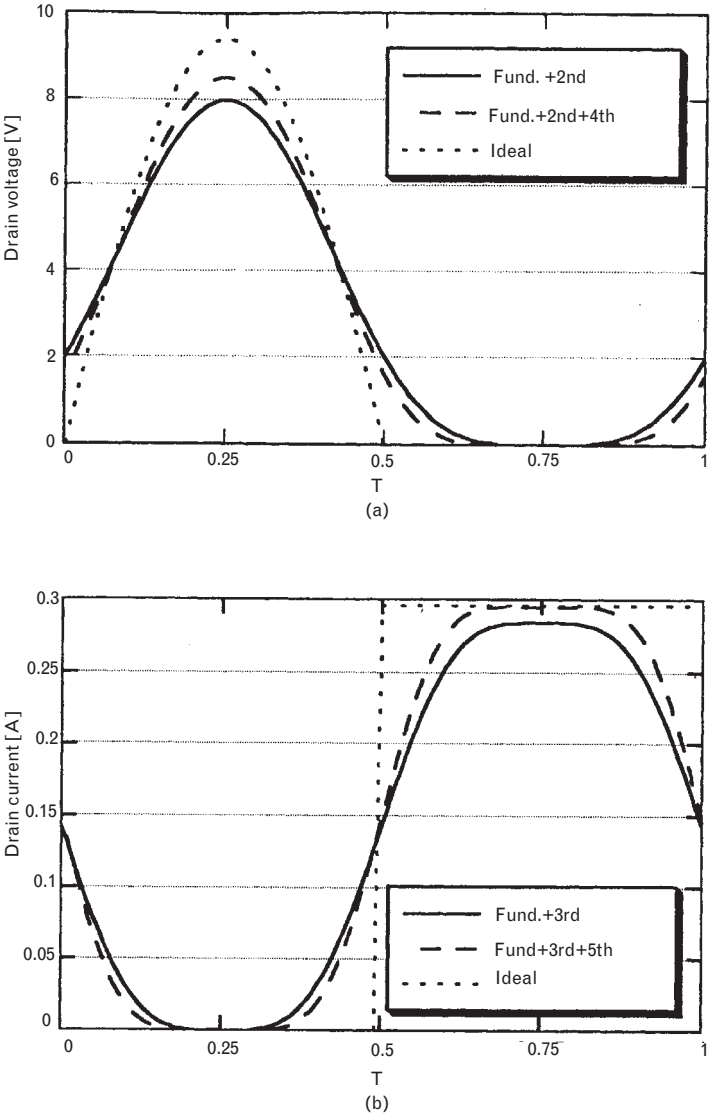


FIGURE 5.37
(a) Voltage and (b) current waveforms for the current-mode class-D switching amplifier, as harmonics up to the fifth are included. (From: [10]. © 2001 IEEE. Used with permission.)

impedance for a square-wave current. The efficiency progressively increases from 75% with three harmonics to the theoretical maximum of 100% as all harmonics are tuned.

5.3.5.2 Class-E amplifiers

The class-E amplifier is similar to the class-B or F amplifier in that the device is (normally, but not necessarily) driven with a 180° conduction angle between saturation and cutoff. However, it uses a switching principle so that any analog relationship between the input and output is lost. The input may be a square wave or a sine wave of sufficient amplitude to drive the device as a switch. Like the class-B amplifier, one of either the switch current or switch voltage is zero at all instants of time, to maximize the efficiency. Its topology is shown in Figure 5.38(a).

Unlike a class-B amplifier, the device is modeled as a switch in parallel with the device output capacitance, which may be supplemented by an additional shunt capacitance. The total drain or collector current is then alternately steered between the device saturation resistance (i.e., the "closed" or "on" switch) and the total transistor output capacitance (represented as all lumped into C_{DS} in the figure). This is quite different from the class-B amplifier, where the device output current source is assumed to have no capacitance and the total drain or collector current is assumed zero during the off half-cycle.

In the class-E amplifier, just prior to turn-on when the switch closes, we force both the magnitude *and* the slope of the device voltage to be zero by tuning of the output load impedance and inclusion of a finite reactive component in the output voltage. The switch current as a consequence rises normally when turned on. It falls to zero abruptly when opened, with the current flowing at that instant being diverted into the shunt capacitance. The zero absolute value and slope of the output voltage when the device is switched on minimizes the energy stored in the device capacitance, which would otherwise be lost as dissipated power. It also relaxes the switch speed requirement. The waveforms no longer require the ideal square-wave behavior and are more analog in nature, much better suited to a high-frequency environment. The inclusion of an output reactance in the load to achieve this can also be useful, since then, in theory, the output capacitance of the device can be absorbed into the load.

With a 50% duty cycle, the *extrinsic* drain load at the fundamental frequency that provides the class-E waveforms is inductive and given by

$$Z_L = \frac{0.28}{2\pi f C_{DS}} e^{j49.05°} \tag{5.41}$$

and is independent of both the input drive level and the drain supply voltage. At microwave frequencies, the nonlinear nature of the intrinsic

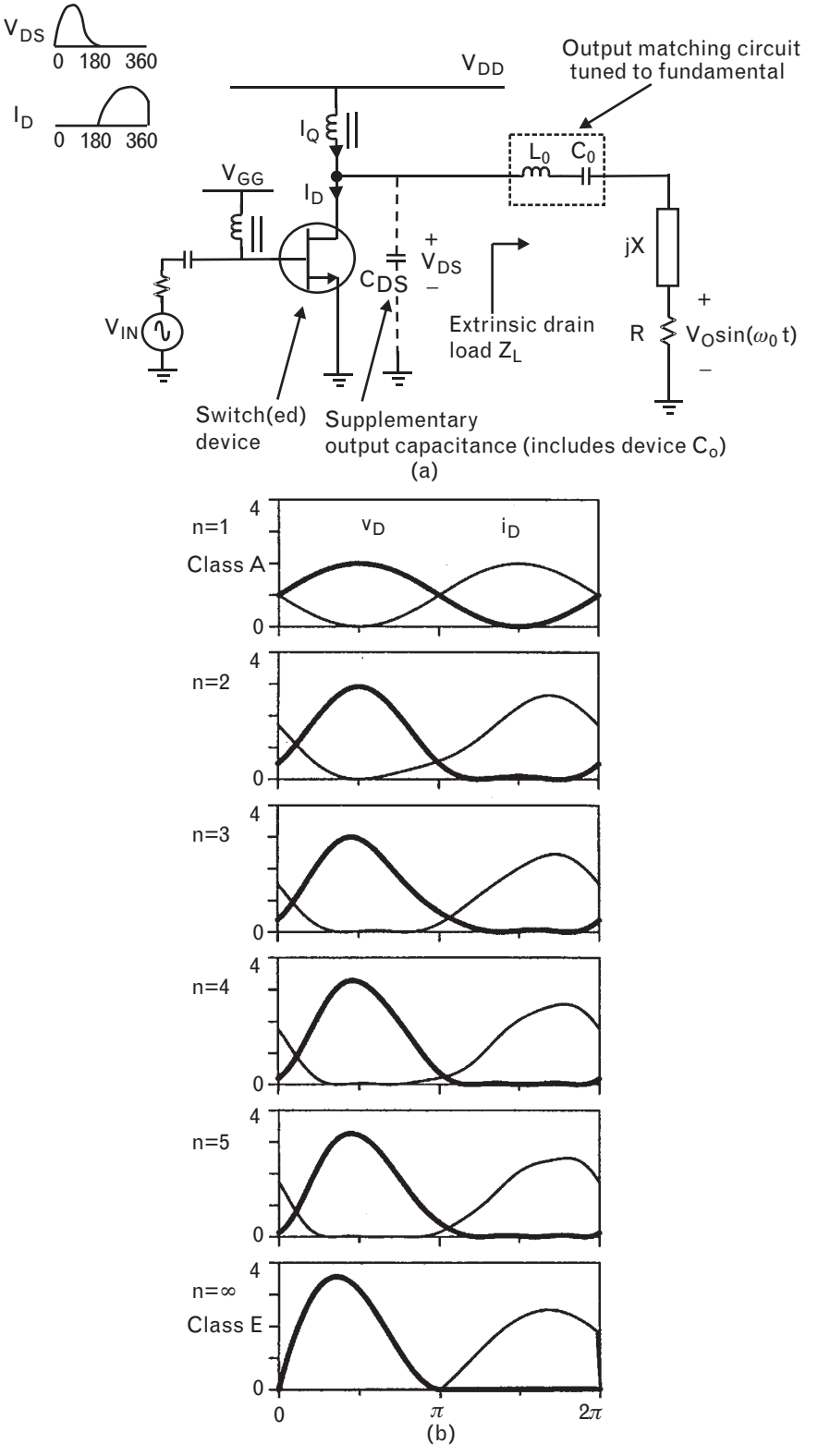


FIGURE 5.38
(a) Topology and (b) waveforms of class-E operation as various numbers of harmonics are included. [Figure 5.38(b) is from [9]. © 2001 IEEE. Used with permission.]

component of C_{DS} and the package parasitics make exact calculation of this optimum load somewhat difficult. Equation (5.41) simplifies to the following sets of approximations for the fundamental load impedance:

$$Z_L = R(1 + j1.1525) \tag{5.42}$$

(which satisfies the 49.05° angle requirement), and

$$\frac{R}{|X_{C_{DS}}|} = 0.1836 \tag{5.43}$$

where $|X_{C_{DS}}|$ is the output reactance of the device itself. The value of R for optimum power is given by

$$R = 0.577 \frac{(V_{DD} - V_{SAT})^2}{P_{RF}} \tag{5.44}$$

Equations (5.43) and (5.44) are not usually simultaneously satisfied at RF because the device output reactance is typically small and the device capacitance may already exceed the requirement calculated. Additional shunt capacitance may be added at the drain if not. However, the efficiency is not unduly sensitive to errors in meeting this constraint, although the power will be suboptimal [12].

Equation (5.42) indicates a key difference between the class-E output and that of most other amplifier classes—that is, the fundamental harmonic reactance is nonzero and noninfinite; rather, it is comparable in magnitude to the fundamental-frequency load resistance. Because of this, the output power from class-E is always less than for class-F [9].

As shown in Figure 5.38(a) the load matching topology to achieve R itself will ideally be implemented as a *series* resonant circuit tuned to the fundamental frequency, in series with the total inductance jX dictated by (5.42). In this way, the load has no *resistive* component at harmonic frequencies, so there is no dissipation there. The point of the series RLC implementation is to make the harmonic terminations high reactive impedances [12, 13] and to keep the voltage and current components in phase quadrature at all harmonics. If only the second-harmonic component termination were accounted for, then the real and imaginary parts of the load impedance would be at 45° (instead of 49.05°) at the fundamental to maintain quadrature at the second harmonic. Raab [9] accounts for the correct harmonic terminating impedances and derives the device voltage and current waveforms of class-E operation in Figure 5.38(b) for increasing numbers of harmonic components. If there are no harmonics ($n = 1$), the

amplifier operates class-A; adding harmonics helps make the waveforms maximally flat and achieve class-E. He also shows that if the second-harmonic reactive load termination progressively increases from a short-circuit to infinite, and the third-harmonic reactive load termination progressively decreases from infinite to zero, then the amplifier changes from class-F operation at one extreme, through class-E, into inverse class-F at the other extreme.

The class-E amplifier has not been widely deployed at microwave frequencies to date, although amplifiers at 900 MHz have been built with output powers approaching 3W and power-added efficiencies around 70% [14]. Apart from its nonlinear nature, one drawback is that the peak voltage swing is 3.56 times the supply voltage, because of the reactive component in the output voltage. Thus, the breakdown voltage needs to be considerably higher for a class-E driven device than for class-F [15].

5.3.6 Cascaded power amplifier design

As illustrated in previous chapters, when there is insufficient gain from a single stage it can be necessary to cascade stages. The same principles apply here. The interstage matching network should ideally present the optimum load impedance to the first device and shape the source impedance appropriately for the second device. Flattening of gain over frequency needs to be achieved within the above constraints.

With power amplifiers, the cascade design needs to be carefully analyzed for both gain and power (and of course stability, which we have taken so far, perhaps unreasonably, for granted). It is quite possible to design a cascade that provides good small-signal gain but has insufficient power to drive the following stage. This is especially so when the matching networks have to introduce reactive mismatch loss at low frequencies to compensate for the gain roll-off of each device and achieve flat gain over frequency. It is also desirable to minimize the VSWR between components that will be interconnected: high VSWR between stages can result in variable group delay, reradiation out the antenna, and lead to losses because of the high (reactive) circulating currents that can result. Using lossy matching networks should be considered in some instances to maintain both VSWR and stability.

The best way to check the power and gain response is to assign a budget for each component of the cascade over frequency. For example, consider the cascaded amplifier in Figure 5.39. If this amplifier needs to operate between 1 and 2 GHz, achieve flat gain of 20 dB, and output power of 30 dBm, it is likely to require two stages of gain. The amplifier also requires good input and output VSWR.

Assume the first FET has a 1-dB compressed power of 23 dBm with 13-dB gain, and the second FET has a 1-dB compressed power of 30 dBm and 7-dB gain, measured at 2 GHz. The design at 2 GHz can achieve the

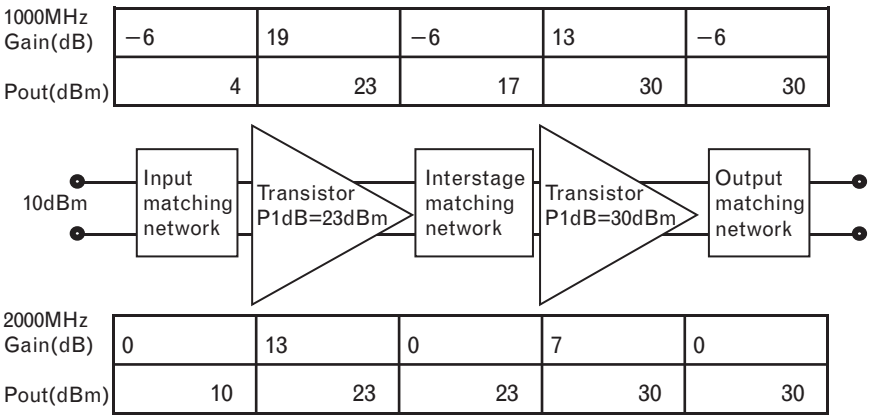


FIGURE 5.39
A cascaded amplifier block diagram, showing the gain and power levels for each stage.

amplifier objectives, assuming that the reactive matching elements used are lossless, since the output matching network of each device can match for the optimum load and the input matching network of each device for conjugate match and maximum power gain.

At 1 GHz, each FET will have approximately 6-dB more gain that at 2 GHz, because the gain of each device rolls off at 6 dB/octave or 20 dB/decade, due to the dominant input pole formed by $R_{IN} - C_{GS}$. But the system requires the same gain as before, 20 dB, and therefore the matching networks need to introduce 12 dB of mismatch loss at 1 GHz.

A major decision is where to distribute this loss. Any loss at the output will destroy output power, which is totally undesirable. Any mismatch loss at the input will reflect power back out the input port and destroy the VSWR. Thus, inserting the full 12 dB of mismatch loss in the interstage matching network achieves both the input VSWR and small-signal gain objective. To maintain interstage VSWR, the loss would best be achieved using a network that diverts excess power at the low frequencies into a resistor. The design of lossy matching networks was covered in Chapter 2.

Unfortunately, inserting 12 dB of loss in the interstage network limits the input power to the final stage to a 1-dB compressed power of 23 dBm – 12 dB = 11 dBm, since the driver compresses at 23 dBm. Even with its 13-dB available gain at 1 GHz, the output stage can only produce an output power of 24 dBm. Although the amplifier meets its specification when driven small-signal, when the input power is increased to the nominal 10 dBm to achieve rated output power, the first stage saturates prematurely. The small-signal gain of the first device at 1 GHz is 19 dB, but its 1-dB compressed output power is still 23 dBm. Thus, with 10-dBm input power, this stage is 6 dB compressed.

We can check the response by calculating the cascaded intercept point introduced in Volume I, Chapter 3. If the intercept point of the first device is 33 dBm (10 dB higher than its 1-dB compression point) and that of the second device is 40 dBm, the cascaded intercept point at 1 GHz is

$$\frac{1}{IP_O} = \frac{1}{33\ \text{dBm} + (-12 + 13)\ \text{dB}} + \frac{1}{40\ \text{dBm}}$$

$$IP_O = 33.03\ \text{dBm}$$

(where the addition is performed in real ratios, not decibels). This is clearly dominated by the first device rather than the second power device. This implies that we are using an expensive power transistor that has negligible impact on the output power.

The only solution, therefore, is to redistribute the 12-dB loss needed for flat gain over frequency. At 1 GHz, we need to put 6-dB mismatch loss at the input and 6 dB in the interstage. As shown in Figure 5.39, the distribution of power throughout the amplifier now meets the specification for both gain and power, as the first device is driven at its rated power, which is necessary to drive the output stage at its rated output power. The two devices are driven at equal levels of power backoff with respect to their 1-dB compression point. It makes intuitive sense to size devices like this, since this avoids unnecessary expense in oversizing the driver device, while minimizing the distortion contribution from either the driver or the output. A better, though more costly, solution is to oversize the earlier stages to maintain an output intercept point close to that of the final stage. Then, the early stages are linear amplifiers and the output stage sets the linearity of the entire chain.

By adding loss at the input in our redistribution of gain we have created a new problem. If the loss is achieved through mismatch (i.e., reactive components), we are reflecting incident power back out the input and destroying the VSWR. The most obvious remedy is to use a resistive matching network at the input rather than a reactive network, so that some of the input power is diverted to a lossy resistor at 1 GHz rather than reflected. This will degrade the noise figure at low frequencies. Another remedy is to use an isolator preceding the reactive matching network to divert the reflected power into a third port that is resistively terminated. This is a good solution, but a broadband isolator adds cost. Another solution would be to modify the first device itself in an attempt to flatten its gain; resistive feedback around the device might be a good way to do this. A fourth solution would be to live with the VSWR problem but use a second identical stage in a balanced configuration to cancel the reflected power. If there is sufficient space, the balanced amplifier is probably the best solution since it brings with it the other advantages outlined in Chapter 2.

5.4 Power amplifier design example

There are numerous ways to design a power amplifier, and as CAD tools become more powerful, the sequence of steps to follow becomes simpler. Today, with load-pull tuning available in many CAD systems, the optimum

load impedances can be automatically found, and the design becomes one of most conveniently synthesizing the linear circuit represented by the tuner impedances. This is the approach we will follow below in the design of a single-stage power amplifier to operate in the 1,900-MHz wireless band. The minimum required output power is 500 mW (1-dB compressed), the battery voltage is 3V, and the required gain is 10 dB.

5.4.1 Transistor selection

There are a large number of devices that can be chosen for designing a power amplifier. The characteristics of many of these devices were reviewed in Chapter 3. In the 1.9-GHz band, the frequency is still low enough that any of the FET variants (including MESFETs, HEMTs, and even MOSFETs), or BJT and HBT variants, will be suitable. However, the current density necessary to achieve 500 mW from a 3-V battery will very quickly eliminate most devices.

For this application, we will assume for now a V_{SAT} of 1V (this turns out later to be a higher—or more conservative—estimate than needed), so that with a battery voltage of 3V the zero-to-peak voltage swing can be 2V. For 500-mW output power, the required zero-peak current swing can be derived from (5.7),

$$I_{PEAK} = \frac{2P_{RF}}{V_{PEAK}} = \frac{1{,}000 \text{ mW}}{2\text{V}} \approx 500 \text{ mA} \tag{5.45}$$

For a class-A amplifier, this is the *minimum* dc current required to support this RF power level, assuming the device input is driven sufficiently hard to swing the current to zero on one-half of the cycle. The typical dc current to support relatively linear operation at this power level will therefore need to be greater than 500 mA, as this will allow higher RF current swings and a larger saturated power output.

At this low a bias voltage, most of the alternative devices will meet the breakdown voltage requirement of 6V in the case of a bipolar and roughly 8 to 10V in the case of a FET (allowing for 1 to 2V negative gate bias). However, the gain specification will eliminate many silicon bipolar transistors, since their f_T may not be high enough to provide the specified gain.

In this application, we will use the NE6500379A GaAs MESFET, first introduced in Section 5.3.2.2. In that application, the FET was biased at pinch-off at the gate to demonstrate class-B operation and used a higher drain voltage of 6V. There, an output power of 34.8 dBm (3W) was ultimately achieved. Now, lowering the drain voltage to 3V will reduce the maximum achievable output power by 3 dB to 31.8 dBm (1.5W) if the quiescent current is unchanged. In fact, the power reduction will be considerably more than 3 dB because of the greater influence of V_{SAT} on the

voltage swing V_{PEAK} at low bias voltages. Typical performance curves at 1,950 MHz for a drain bias of 3V are shown in Figure 5.40.

From the left figure, we see that the measured 1-dB compression point with a class-A quiescent current of 600 mA occurs at an output power of 28.5 dBm. From the right-hand figure, we see that with a quiescent current of 800 mA, a saturated output power around 31 dBm can be achieved across the 1.9- to 2-GHz band, with a small-signal gain of over 11 dB. When biased class-AB at a small-signal bias current of 100 mA, the saturated output power is about 30 dBm and the gain is lower. In the latter case however, the dc current will increase strongly as the input drive is increased, supporting the necessary signal swings for the output power. However, if linearity is an important consideration in this amplifier, then a class-A approach should be taken, using a quiescent current somewhat greater than 500 mA. The amplifier efficiency could then be improved using harmonic tuning. We will develop this design at 800-mA bias current to enable comparison with the measured data, and to adequately allow for a peak current swing of 500 mA at the 1-dB compression point. If desired, the bias voltage could be adjusted later to yield a lower small-signal quiescent current (class-AB), and the design reoptimized at the new bias conditions.

5.4.2 Transistor characterization

The bias network for a MESFET is straightforward. The I-V curves (which can be seen in Section 5.4.3) dictate a gate voltage of around –1.75V to achieve the desired 800-mA drain current. This is achieved with a simple

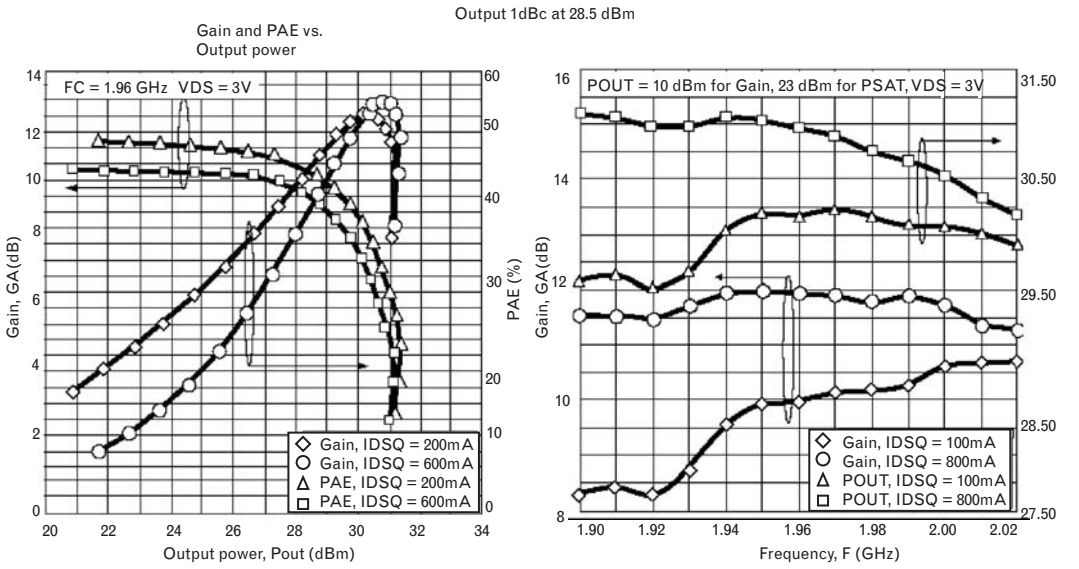

Output 1dBc at 28.5 dBm
Gain and PAE vs.
Output power
FC = 1.96 GHz VDS = 3V
Gain, GA(dB)
PAE (%)
Output power, Pout (dBm)
Gain, IDSQ = 200mA
Gain, IDSQ = 600mA
PAE, IDSQ = 200mA
PAE, IDSQ = 600mA
POUT = 10 dBm for Gain, 23 dBm for PSAT, VDS = 3V
Gain, GA(dB)
Frequency, F (GHz)
Gain, IDSQ = 100mA
Gain, IDSQ = 800mA
POUT, IDSQ = 100mA
POUT, IDSQ = 800mA

FIGURE 5.40 *Measured gain and power-added efficiency versus output power, and the gain and saturated output power versus frequency, at a drain bias of 3V, for the NE6500379A MESFET. (Courtesy California Eastern Labs.)*

dc voltage source. A 20-Ω resistor in series represents the internal impedance of the voltage source and also allows for placement of a small chip resistor at the gate terminal to improve bias-network stability and prevent oscillation through the bias network.

The optimum load line to force the device to swing between zero and the supply rail voltage of 3V and with a zero-peak current of 500 mA is given from (5.6). It is simply the inverse slope of the load line, defined by its endpoints as

$$R_{OL} = \frac{2\left(V_{DD} - V_{SAT}\right)}{I_{MAX}} = \frac{\left(V_{DD} - V_{SAT}\right)}{I_{PEAK}} = \frac{\left(3-1\right)}{0.5} = 4\Omega \qquad (5.46)$$

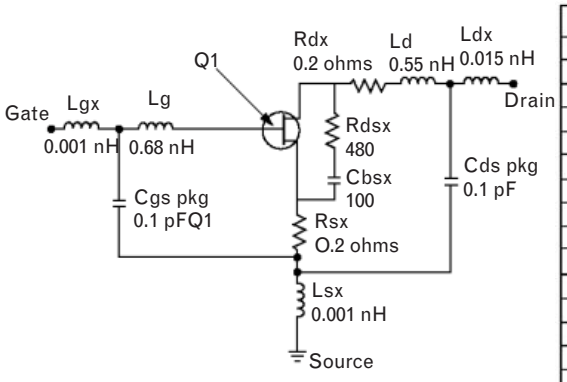
This is the load resistance that needs to be impressed at the intrinsic terminals of the MESFET drain current source. It will be a different value at the extrinsic (external) terminals because of transformation through the package parasitics, and feedback through the device itself. The model used to simulate the MESFET is shown in Figure 5.41.

The device consists of a NEC65003 chip in a surface-mounted 79-A package. The package parasitics are shown, and as we will see shortly, are important as they enable the device to be "de-embedded" correctly. The nonlinear model enables verification of the dc bias conditions, calculation of the small-signal *S*-parameters at the chosen bias point, and simulation of the large-signal performance of the device.

Once the small-signal *S*-parameters of the device are calculated, then the result of (5.46) can be used as a first approximation to determine the extrinsic load impedance at the drain and how the amplifier should be matched. The intrinsic optimum load resistance of 4Ω is the resistance that must be impressed at the internal drain current source. That current source must support the required voltage swing between ground and the supply rail. If the output-matching network transforms the external 50-Ω load resistor to 4Ω at the terminals of the current source, then the same matching network will transform a 4-Ω resistor placed at these terminals to 50Ω externally.

Figure 5.42 shows the principle. The calculated small-signal *S*-parameters of the device were used to derive the linear equivalent circuit, in which the input of the FET is modeled as a series R-C network, the output of the FET as a current source with a finite output impedance shunted by a capacitor, and a feedback capacitor between the input and output. With the exception of the current source, these elements are shown within the box of Figure 5.42, embedded within the device parsitics[1] taken from Figure 5.41. However, the small-signal current source has been replaced by the optimum load resistor, 4Ω. It represents the optimum impedance

1. The series RC combination of RDBX and CBSX at the output of the transistor in Figure 5.41 are not parasitics but model the dispersion, or frequency dependence, of its output resistance.



Fet nonlinear model parameter(1)

Parameters	Q1	Parameters	Q1
VTO	−2.585	RG	0.2
VTOSC	0	RD	0.001
ALPHA	3	RS	0.001
BETA	1.28	RGMET	0
GAMMA	0	KF	0
GAMMADC(2)	0.035	AF	1
Q	1.7	TNOM	27
DELTA	0.02	XTI	3
VBI	0.6	EG	1.43
IS	1e-14	VTOTC	0
N	1	BETATCE	0
RIS	0	FFE	1
RID	0		
TAU	10e-12		
CDS	0.5e-12		
RDB	0.001		
CBS	0		
CGSO(3)	25e-12		
CGDO(4)	4.5e-12		
DELTA1	1		
DELTA2	0.2		
FC	0.5		
VBR	Infinity		

(1) Series IV libraTOM model

FIGURE 5.41 *The schematic of the NE6500379A MESFET and the parameters for the nonlinear model. (Courtesy California Eastern Labs.)*

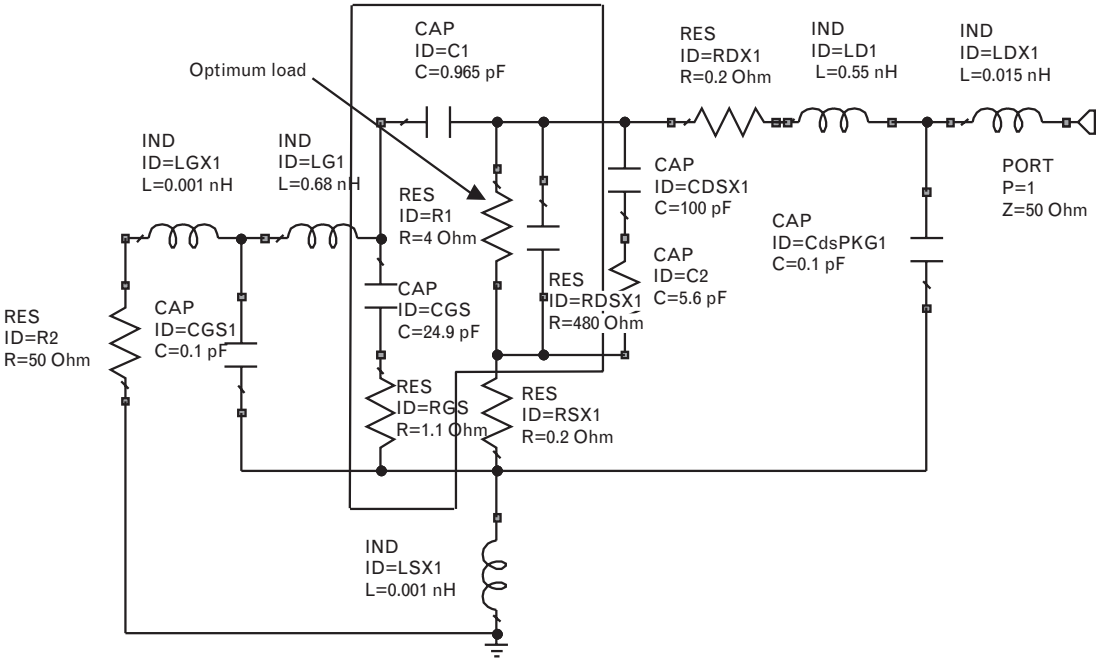


FIGURE 5.42 *The linear MESFET model, with the optimum load resistor of 4Ω replacing the voltage controlled current source.*

the external load should impose, and thus see, at these terminals, under large-signal excitation.

The impedance seen looking into the extrinsic drain terminal is shown in Figure 5.43. This is just the conjugate of the optimum load impedance that should be placed at the external drain terminal in order for the device current source to see 4Ω. The current source can then generate maximum voltage and current swing, hence optimum RF power. If there are no losses within the device or matching network, then this is the power transferred to the load. From the figure, we see that the 4-Ω resistance has been transformed at the external drain terminal by the feedback of the device, and by the package parasitics, into an impedance of $4+j6\Omega$. Thus the device should be terminated with a load of $4-j6\Omega$ at its drain.

This is an extremely simplistic approach since it neglects the nonlinear behavior of the device at large-signal levels. However, it is always easiest to begin with the quasi-linear approach in order to more easily synthesize a starting point in designing the matching networks.

The validity of this approach can be validated by examining load pull data. The load pull contours of the FET are shown in Figure 5.44, simulated using the full nonlinear FET model. A point on the 33-dBm contour is shown, with (unnormalized) impedance $3.2-j5.5\Omega$. The central contour is for a saturated output power of 33.5 dBm, and the contours are in 0.5-dB

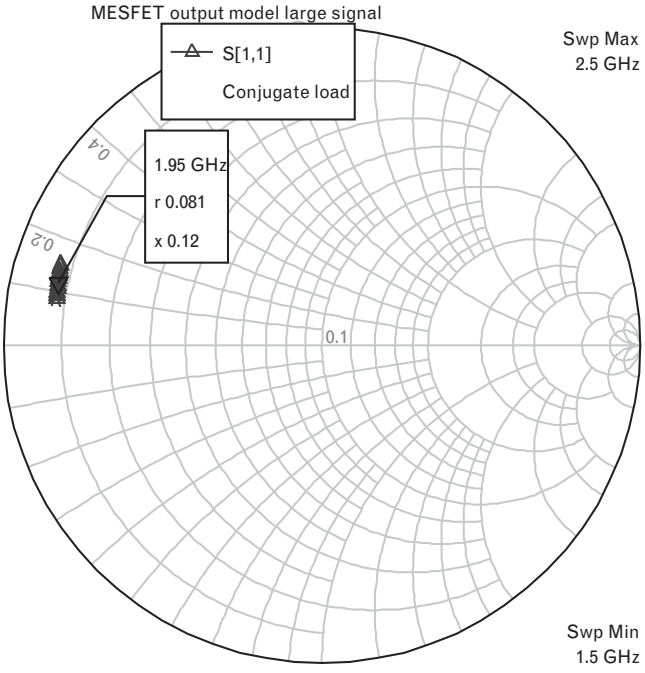


FIGURE 5.43 *The impedance looking into the extrinsic drain terminal with the optimum load resistor at the intrinsic drain terminals replacing the drain current generator.*

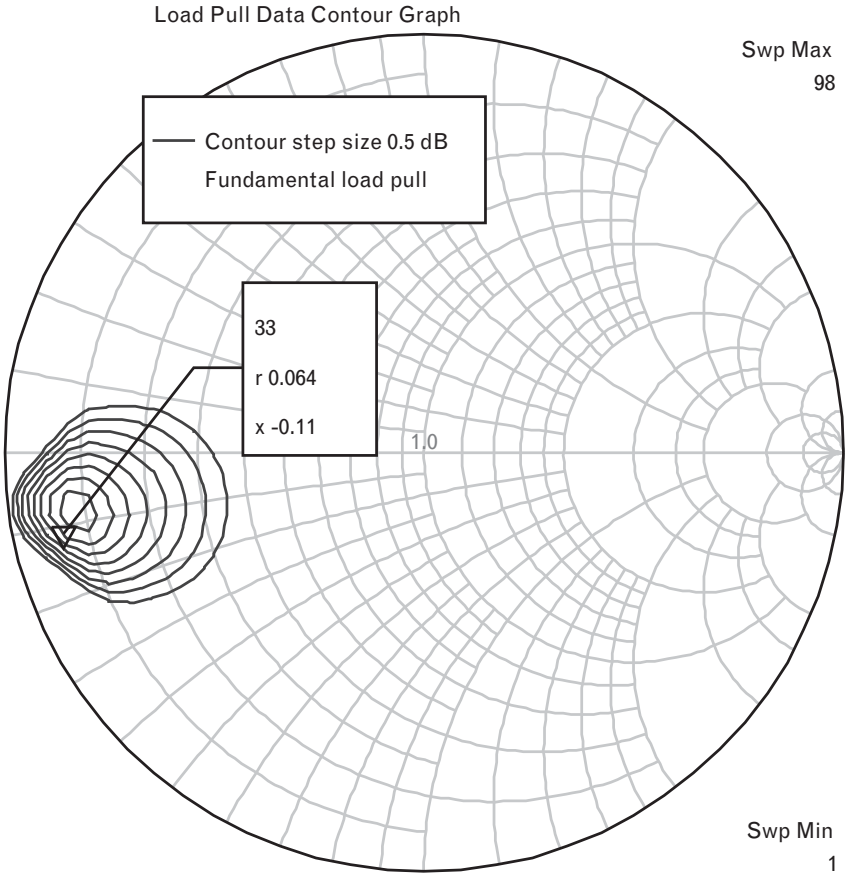


FIGURE 5.44
Simulated load pull contours for the MESFET. The center contour has 33.5 dBm saturated output power. Drain voltage is 3V and drain current 800 mA.

steps. Although this level of saturated output power is unreasonably optimistic, the contours do verify that the optimum load impedance at the (extrinsic) drain is very close to the value we predicted above from quasi-linear considerations, $4-j6\Omega$. The higher predicted power results from a full 800-mA current swing at the drain, rather than our assumed 500 mA.

The simplistic quasi-linear approach is reasonably accurate in this case for two reasons. First, the package parasitics, and in particular C_{DS}, often tend to dominate the impedance seen at the external device terminals and can mask small errors in the choice of optimum load resistor. Second, the load resistance is so low that it tends to swamp out other errors. Unfortunately, as we will shortly see, these two effects also work against us and can make the design of power amplifiers very susceptible to tuning error.

5.4.3 Matching the input and output of the device

As we have just noted, the center of the nonlinear load pull simulations gives the impedance for maximum output power, while impedance values that achieve lower values of compressed power lie around it. Using the

load pull contours, we select load impedances of 1.25–*j*5Ω on the gate and 3.2–*j*5Ω on the drain (at 1,950 MHz) for the input and output match, respectively, to achieve 1-dB power compression.

The amplifier problem is then reduced to "simply" synthesizing these values of impedances with either lumped or distributed elements. The tuners have provided a simple way of first characterizing the necessary impedances and then modeling the device behavior with input drive, bias, and changes in harmonic terminations.

The amplifier characteristics when terminated with these impedances are plotted in Figure 5.45. The simulated small-signal gain is 16.4 dB, and the 1-dB compressed output power is 29.2 dBm. Saturated output power is more than 32 dBm at these conditions and consistent with the load pull contours. The quiescent current remains fairly constant around 800 mA with input power because the amplifier is biased class-A.

Figure 5.46 shows the simulated output load line of the device when driven at 1-dB compressed power. At the intrinsic terminals of the device, the current and voltage are in phase, but at the extrinsic terminals the load line is elliptical. There, the voltage is phase-shifted from the current, due to the effect of the parasitics in transforming the impedance to a new

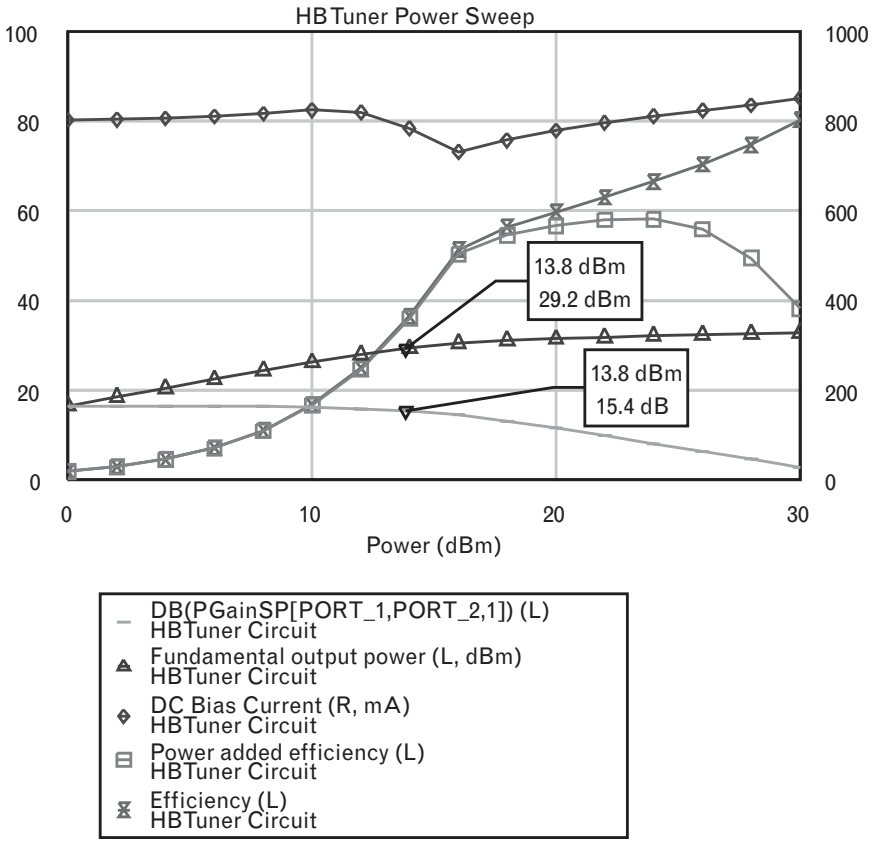


FIGURE 5.45
Simulated amplifier characteristics when the MESFET is tuned with the selected fundamental input and output impedances by load pull tuners. Higher harmonics are terminated in a short circuit. The bias current (milliampere, right axis) and behavior of output power, gain, efficiency, and power-added efficiency (left axis) are shown versus input power at 1,950 MHz.

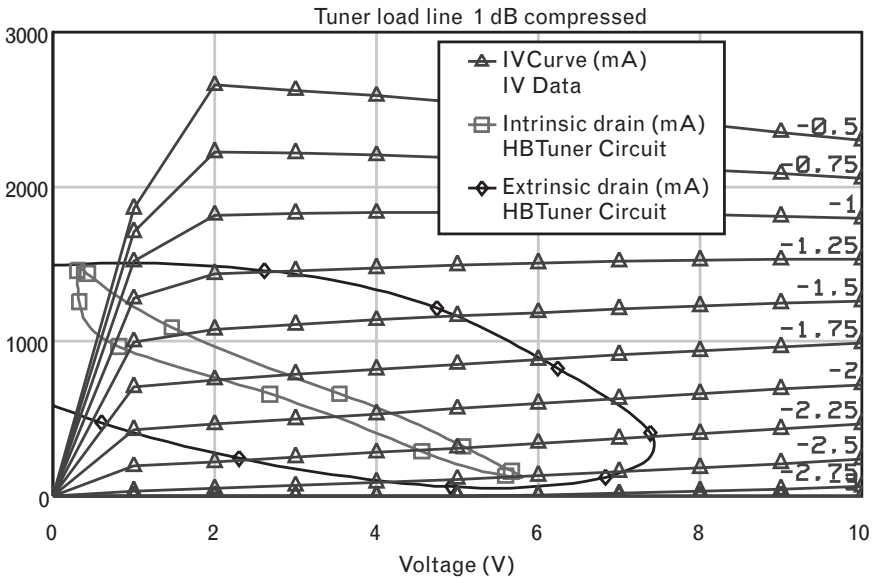


FIGURE 5.46 *Simulated load line characteristics at the 1-dB compression point when the device is tuned with the selected fundamental input and output impedances by tuners. Higher harmonics are terminated in a short circuit.*

reference plane. This shows the importance of properly accounting for device parasitics at RF and higher frequencies. The intrinsic drain voltage is clamped at V_{SAT} (the simulated value is around 0.3V) when the gate voltage swings up to zero or just slightly above it, and on the negative swing reaches $2V_{DD}$ (6V). The extrinsic device behavior is quite different, with the extrinsic drain voltage even swinging negative over portion of the cycle. This effect is due to the quite large (0.55 nH) series drain inductance in the package model, which is resonated by the high-Q output load presented by the tuner.

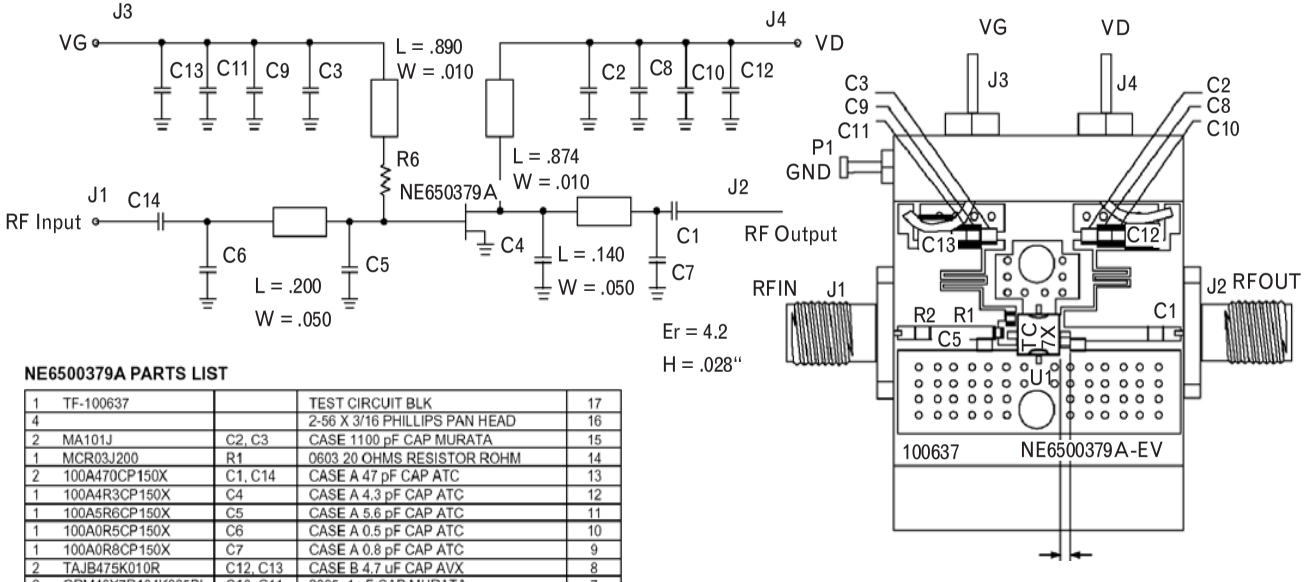
We next compare our design and simulated results with a measured test amplifier, whose schematic and layout (taken from the data sheet) are shown in Figure 5.47.

The input and output matching networks have been realized using lowpass shunt C–transmission line–shunt C sections. The transmission lines are fabricated as microstrip and should provide terminating impedances close to those we have modeled using the tuners. We will verify this shortly. The circuit used for initial simulation reflects the physical layout and is shown in Figure 5.48.

The swept power and load line characteristics of Figures 5.49 and 5.50 result.

The small-signal gain is 12 dB, and the 1-dB compressed output power is simulated to be 28 dBm. The load line characteristic at the 1-dB point and the load impedance in Figure 5.50 appear quite similar to those in Figure 5.46 we simulated in our initial design, with the load pull tuners.

It is sometimes helpful to reconstruct the waveforms of the voltage and current at the intrinsic device terminals, to visualize the load line in the



NE6500379A PARTS LIST

1	TF-100637		TEST CIRCUIT BLK	17
4			2-56 X 3/16 PHILLIPS PAN HEAD	16
2	MA101J	C2, C3	CASE 1100 pF CAP MURATA	15
1	MCR03J200	R1	0603 20 OHMS RESISTOR ROHM	14
2	100A470CP150X	C1, C14	CASE A 47 pF CAP ATC	13
1	100A4R3CP150X	C4	CASE A 4.3 pF CAP ATC	12
1	100A5R6CP150X	C5	CASE A 5.6 pF CAP ATC	11
1	100A0R5CP150X	C6	CASE A 0.5 pF CAP ATC	10
1	100A0R8CP150X	C7	CASE A 0.8 pF CAP ATC	9
2	TAJB475K010R	C12, C13	CASE B 4.7 uF CAP AVX	8
2	GRM40X7R104K025BL	C10, C11	0805 .1 uF CAP MURATA	7
2	GRM40C0G102J050BD	C8, C9	0805 1000 pF CAP MURATA	6
1	NE6500379A	U1	IC NEC	5
1	703401	P1	GROUND LUG CONCORD	4
1	1250-003	J3, J4	FEEDTHRU MURATA	3
2	2052-5636-02	J1, J2	FLANGE MOUNT JACK RECEPTACLE	2

FIGURE 5.47 *The data sheet schematic and layout of a 1,950-MHz amplifier using the NE6500379A. (Courtesy California Eastern Labs.)*

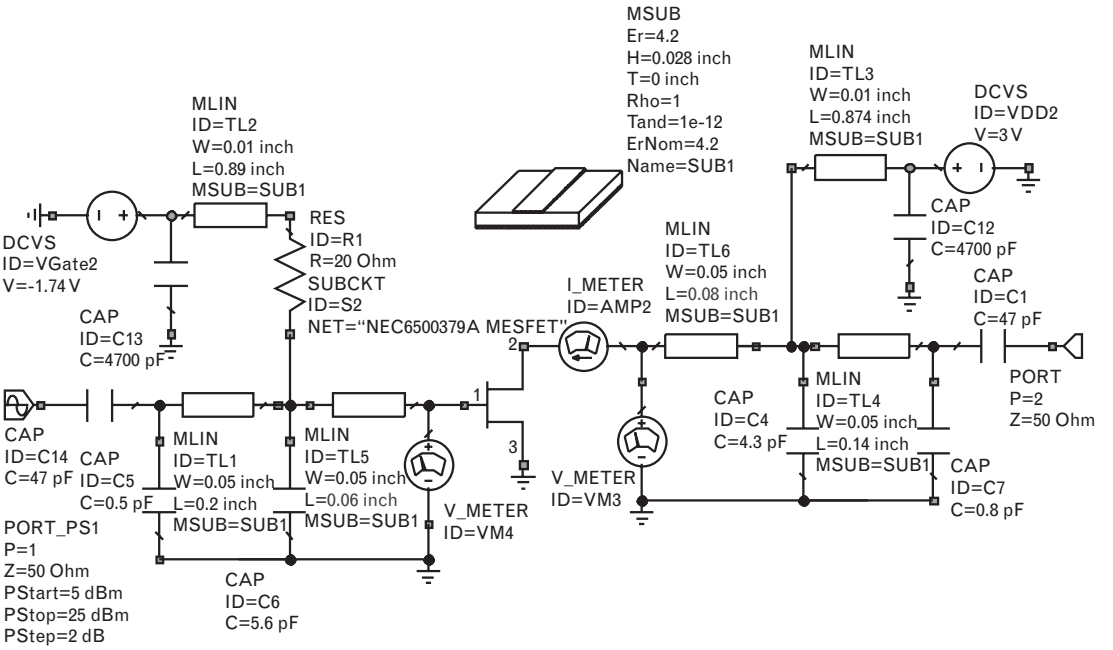


FIGURE 5.48 *Amplifier schematic of Figure 5.47, reflecting the physical layout of the test amplifier.*

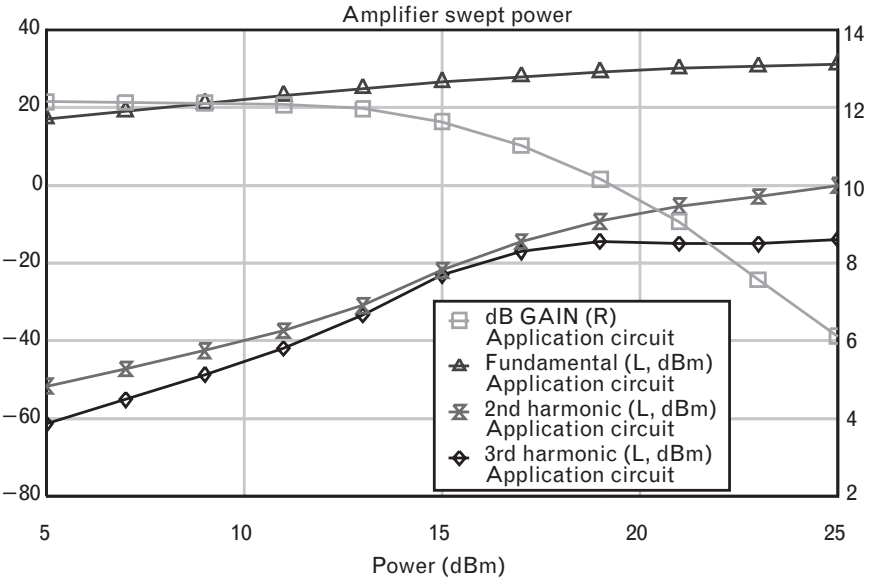


FIGURE 5.49 *Simulated swept power characteristics of the MESFET amplifier of Figure 5.48.*

dimension of time. These are shown in Figure 5.51 and show the effect of the device package on the drain voltage. At the extrinsic reference plane, the drain voltage is shifted by almost 90° when compared with the intrinsic drain voltage. If this were not accounted for by the matching network,

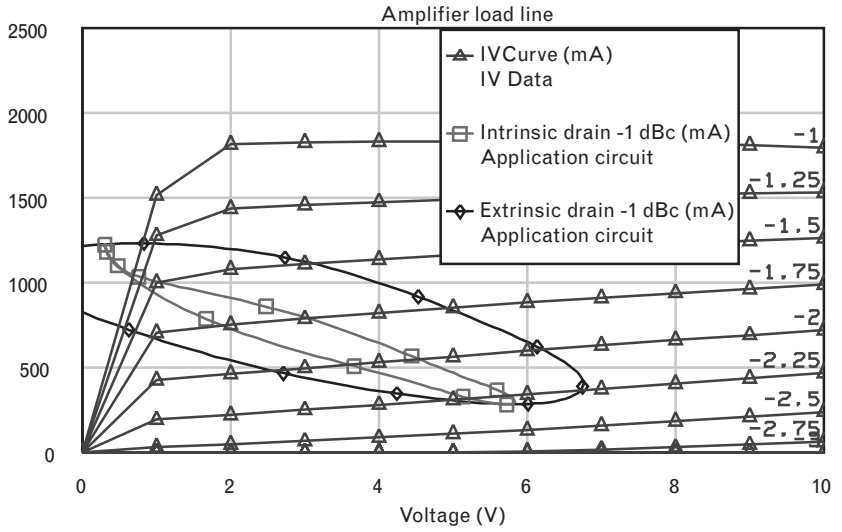


FIGURE 5.50
Simulated load line characteristics of the MESFET amplifier of Figure 5.48.

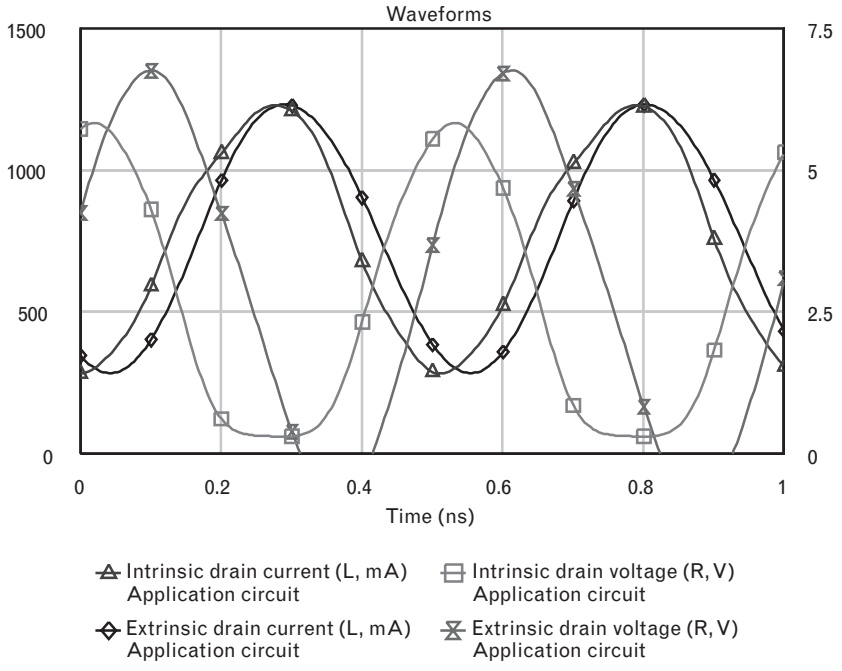


FIGURE 5.51
Waveforms of the intrinsic drain current and voltage, and extrinsic drain current and voltage of the MESFET amplifier of Figure 5.48.

which restores the desired in-phase relationship at the intrinsic terminals, most of the device output power would be reactive.

As illustrated by the differences between the intrinsic and extrinsic simulations, the effect of the device package is critical in the matching scheme, and even small errors can cause very large transformations in the load match. In fact, the layout of Figure 5.47 contains small transmission

lines right at the gate and drain terminals that are not shown in the schematic in that figure. Close examination of the layout shows that between the input and output of the FET, and the first shunt capacitances, there are small sections of transmission line. These small lengths of a 50-Ω line appear negligible, but they were needed above to properly model the matching circuits. We estimated their lengths as 0.060 inch (1.5 mm) at the input and 0.080 inch (2 mm) at the output in simulating the results above. If we omit these transmission lines from the simulation, then the performance shown in Figure 5.52 results, which is significantly worse.

This circuit would now achieve only 5.7-dB small-signal gain and a 1-dB compressed output power of 25 dBm—well under specification. The reasons for the poor power performance can best be seen on the simulated load line characteristic, plotted in Figure 5.53 at the 1-dB compression point.

Several problems are evident from the intrinsic load line. First, the intrinsic drain voltage swing is the same as that of Figure 5.50, but the required input power to achieve it is over 6 dB higher. Second, the peak-to-peak drain current swing is only 900 mA compared with 1,400 mA when properly tuned (Figure 5.46), indicating that the output load

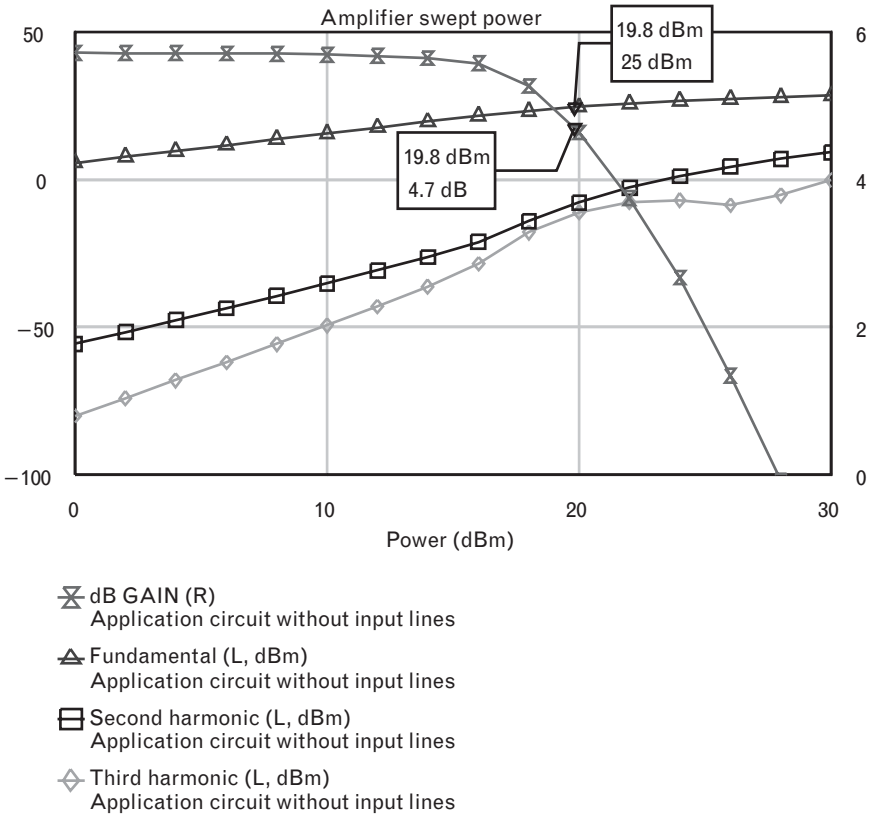


FIGURE 5.52
Simulated performance of the amplifier layout of Figure 5.47 omitting the small line sections at the gate and drain, showing fundamental, second, and third-harmonic output powers, and the fundamental gain, versus input power at 1,950 MHz.

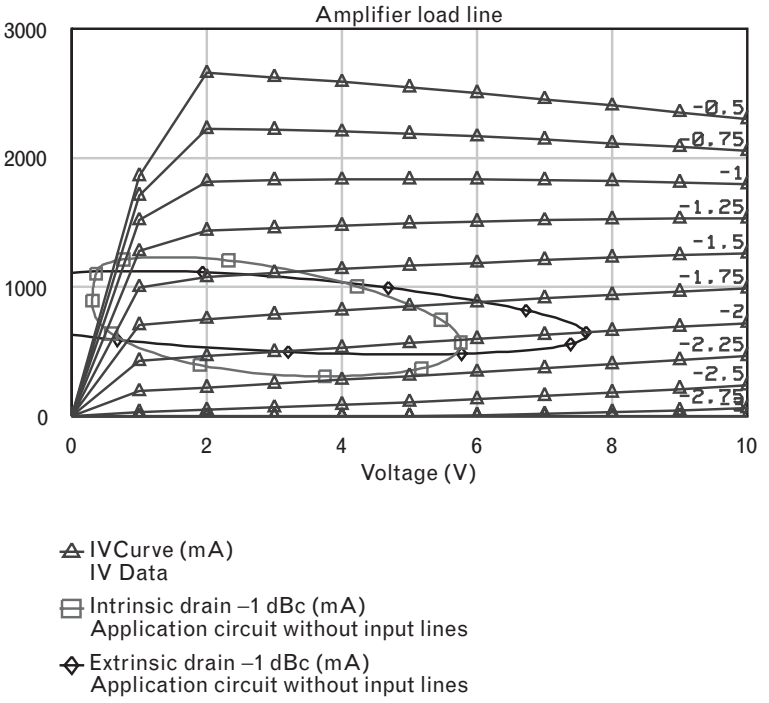


FIGURE 5.53
Simulated load line characteristic of the amplifier layout of Figure 5.47 omitting the small line sections at the gate and drain, at the 1-dB compression point.

impedance is too high. The smaller slope of the load line reflects this. Finally, the intrinsic load line is now elliptical, indicating that the voltage and current are out of phase because of a mismatch presented to the intrinsic drain.

The large differences between these two sets of simulations are due to very small errors in modeling the layout right at the device terminals. The small line lengths included in the first simulation properly represent the layout and add series inductance to the gate or drain. This can inadvertently arise in real circuits in other ways, for instance, due to device variation or improper mounting in a circuit.

How can such a minute change to the circuit make such a drastic change in output power and compression behavior? Figure 5.54 shows how the synthesized input and output matching networks of the built amplifier match the desired response derived from the load pull tuners. Although the matching circuits are not perfect in Figure 5.54(a), they are not far off. However, because the load pull contours are so closely spaced around the optimum load, the sensitivity to load impedance is enormous: a small error in realizing the correct impedance can cause large errors in output power at the load. With the short transmission line sections included at the gate and drain ends of the matching network, the agreement in Figure 5.54(b) is almost perfect and the amplifier meets the required specifications.

This highlights one of the most difficult problems that face the designer of power amplifiers, the need to transform from very low impedance levels

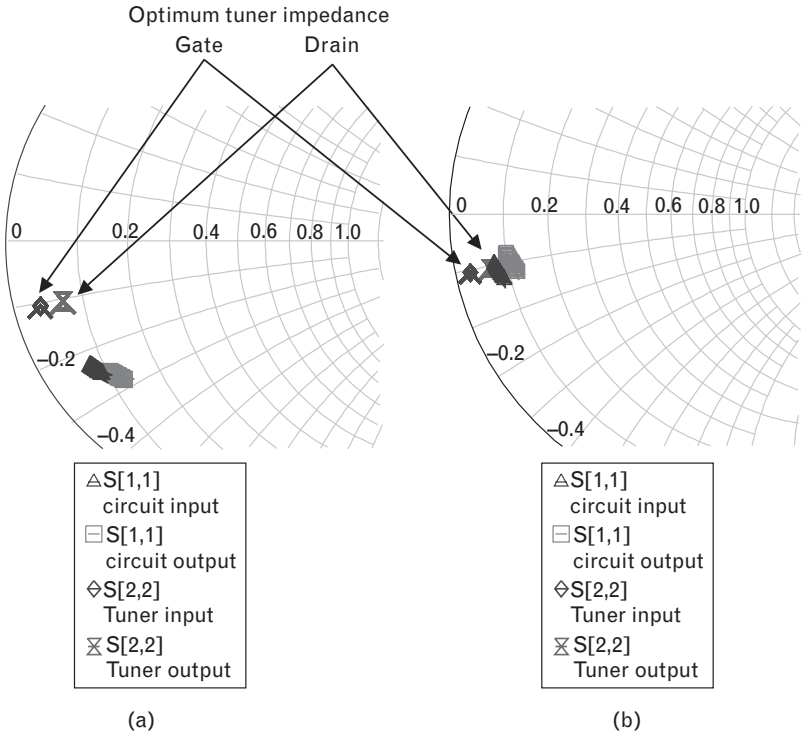


FIGURE 5.54
(a) Impedance of the input and output matching networks of the 1,950-MHz test amplifier without the small gate and drain transmission line sections, compared with the optimum impedance derived from the load-pull tuner; and (b) the impedance of the matching networks, with the small transmission line sections included.

up to 50Ω. The large currents in power devices, and the small battery voltages available, dictate low impedance levels at the device intrinsic terminals. The output capacitance of the device transforms this low dc resistance to even lower values at RF frequencies. Small variations in impedance at the device terminals—of the order of an ohm—can translate into errors of many dBm in power. We have seen that the problem is compounded because the package effects can transform the impedance at the intrinsic device terminals into almost unrecognizable values at the external device terminals. Lack of an accurate package model can make de-embedding of device behavior almost impossible.

Unfortunately, there is no easy work-around. Using multiple smaller devices with higher output impedances and then external power combining is one approach. In some systems, impedance levels other than 50Ω are also used. This is possible in interstage matching networks, where rather than transforming a low-output device impedance to 50Ω, it can be transformed directly to the (low) input impedance of the following stage. Conceptually, this is similar to the approach used in monolithic integration, where impedance transformation is avoided altogether between stages because the physical distances between them are so low that reflections do not arise.

When we consider the design of broadband matching networks to achieve broadband power amplifiers, the transformation of low impedances to 50Ω becomes even more difficult. Even when unpackaged chip

devices are used, the impedance levels remain low, although the bandwidth achieved can be greater because of the absence of package parasitics. Multistage matching network can also help mitigate the problem by broadening the bandwidth of the match and loosening the tolerances on the matching network, but ultimately some tuning might still be necessary. In a two-section matching network, the approach taken is to use the first section to match to an intermediate impedance value (usually the geometric mean of the device and the final system impedance), perhaps at the highest frequency; and to use the second section to complete the match at the lowest frequency.

Some packaged devices are internally matched. The manufacturer places matching elements within the package, close to the chip. This has the advantage that the transformation of impedance levels can be done before the intrinsic bandwidth of the device has been destroyed by the package parasitics and the input leads or transmission lines. These parasitics "stretch" the impedance locus at higher frequencies because of their electrical length. If bandwidths of greater than about 15% are required, then an unpackaged device must be used since the package stretches the impedance to the point that it is impossible to match even reasonable impedance levels over a broader band, let alone low impedances where high-Q transformations are required.

It remains to check the stability of this power amplifier. Because the input and output reflection coefficients of a power device can approach unity, its stability factor (k) is frequently less than one and the device is potentially unstable. The large reflection coefficients also make the device less unilateral even for small reverse gains s_{12}, because of the large magnitude of the reflected waves at both ports.

Indeed, the stability circles of this FET shown in Figure 5.55(a) indicate that the device is potentially unstable when terminated with low impedances at the input and output—the very impedances required for maximum output power. Because our choices for optimum termination impedances lie so close to the boundaries of the stability circles, the input and output match of the amplifier will be poor [since we recall from Chapter 1 that the boundaries of the stability circles themselves indicate the terminations that give the opposite port borderline stability (i.e., a unity reflection coefficient)].

Unfortunately, as shown in the operating gain example using this device in Chapter 2, the only way to push these stability circles outside the Smith chart is to add a small series resistor at either the input or output. Because the stability circles lie at the low impedance side of the Smith chart, adding shunt resistance will not help. Although a series resistance of just 3Ω at the gate will, in fact, achieve unconditional stability, the loss in gain is several decibels, and such a step may be unacceptable to the power designer. The layout of Figure 5.47, in fact, shows a small chip resistor *R2* in series with the gate, precisely to achieve this, although we have ignored

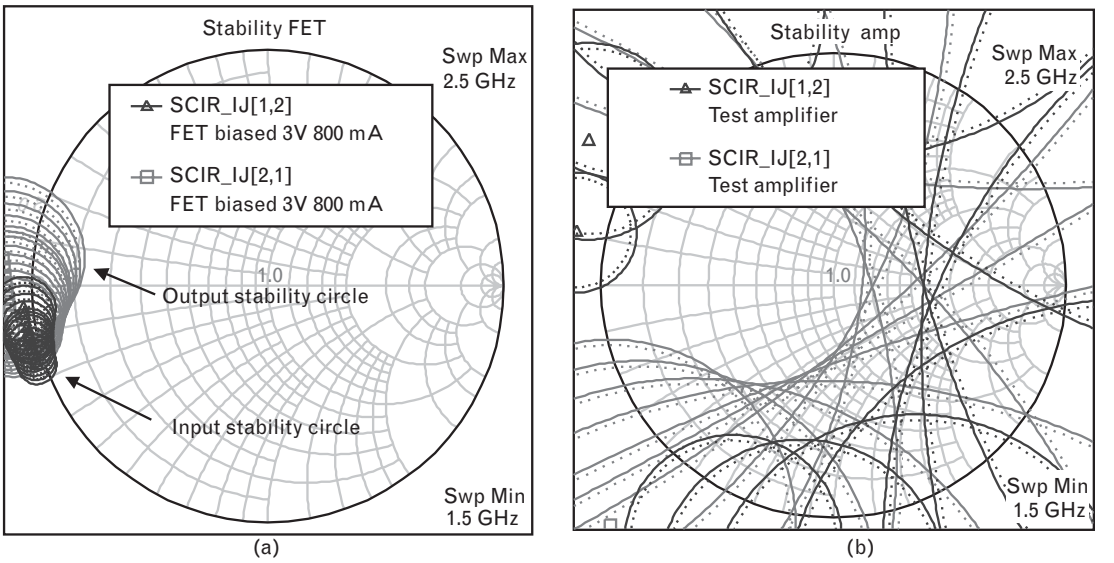


FIGURE 5.55 *The input and output stability circles of (a) the FET and (b) the amplifier shown in Figure 5.48, simulated from 1,500 to 2,500 MHz.*

its presence until now. The stabilization network designed in the example in Section 2.2.4 uses an L-C network tuned to 1.95 GHz in series with such a 3-Ω stabilization resistor, shunted by a 20-Ω resistor to introduce additional loss at other frequencies, and will ensure stability at out-of-band frequencies as well.

The stability circles of Figure 5.55(b) for the amplifier of Figure 5.48 show that as long as the input and output are terminated in 50Ω, the amplifier is (just) stable. The designer can therefore choose between unconditional stability with reduced gain, or conditional stability. In practice, stability will probably be somewhat improved by losses in the dielectric and from radiation that we have not modeled. This is one instance where such losses work at least partially in our favor. The old adage that "you can ship this amplifier with your thumb (which stops it from oscillating)" is well illustrated by cases of marginal stability such as this. However, the design would be marginal at best, given the changing load impedance presented by a mobile handset antenna, and other temperature and production effects.

5.4.4 Harmonic tuning example

As described for class-F operation, if we can force the drain voltage to maintain a low second-harmonic and high third-harmonic component, we can reduce the power dissipated in the device and improve the total efficiency.

Using the load pull tuners at the input and output allows us to test the device with arbitrary harmonic terminations. So far, all the simulations were made with the second and third-harmonic impedances short-circuited at the extrinsic reference planes of the device.

Although the class-F approach of tuning the drain voltage waveform to become square is derived with a device driven class-B, it can still be of benefit to a class-A device when the current is driven between its fundamental limits (of saturation and pinch-off), so that the overdriven class-A device is effectively switched on and off for half a cycle in much the same way as for a class-B device. As we will see in the next section on bias, at high input power levels the quiescent bias point of the class-B device shifts to approach that of the class-A device, so operation when the device is overdriven becomes similar.

Figure 5.56 shows the voltage and current waveforms at the intrinsic drain terminal when the device is overdriven. The device current is driven between the two extremes of the load line, so the overdriven class-A device switches between zero current and 1,600 mA (twice the quiescent current). Keeping the fundamental load impedance constant, the second and third-harmonic components of the output load pull tuner may be varied in order to square up the drain voltage. The response shown is when the second-harmonic impedance is a short circuit and the third-harmonic impedance is an open circuit. Clearly, when the current is high the voltage is almost zero for most of the half-cycle, and the power dissipated in the device is then minimized.

Figure 5.57 shows the simulated response of this harmonically tuned amplifier. The simulation is identical to that shown in Figure 5.45 but with the differing harmonic load impedances. As expected, the 1-dB

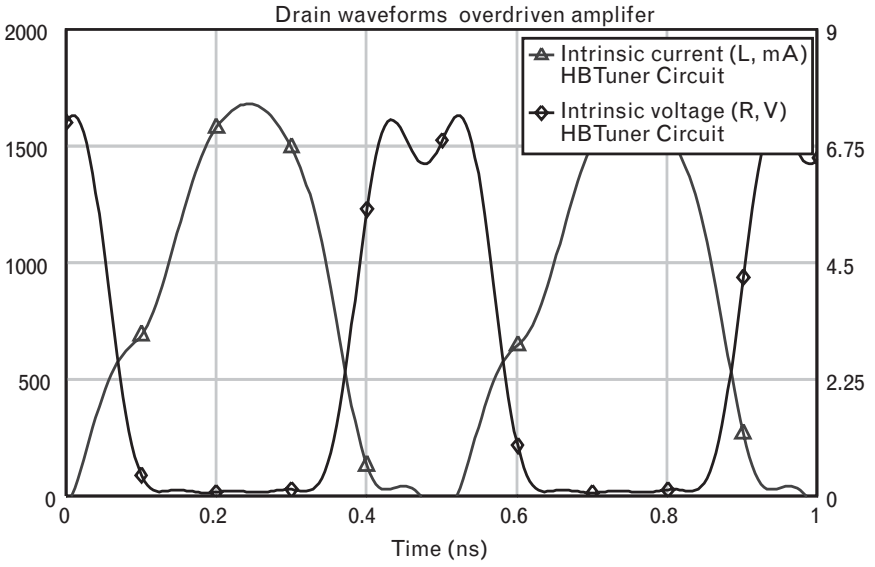


FIGURE 5.56
Intrinsic drain voltage and current waveforms of the MESFET amplifier with harmonic tuning, when driven 6 dB into compression.

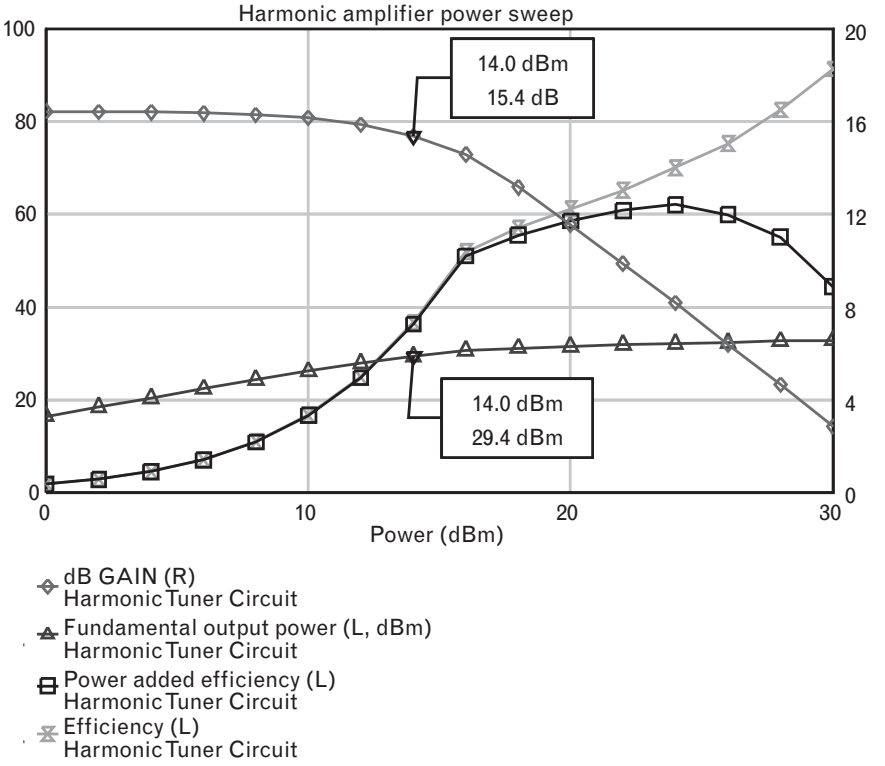


FIGURE 5.57
The simulated gain, fundamental output power, efficiency, and power-added efficiency of the harmonically tuned MESFET amplifier versus input power at 1,950 MHz.

compressed output power is increased slightly—to 29.4 dBm—by squaring up the voltage waveform. If we could do this perfectly, the increase would be even greater (more than 1 dB) because the fundamental component of a square wave is $4/\pi$, or 1.27 times the zero-to-peak value of the square wave itself or of a sinusoid with the same limits. The efficiency is also improved as the device is driven into saturation, since beyond the 1-dB compression point the device is beginning to behave more like a switching amplifier than a linear class-A device.

5.5 Bias considerations

Biasing of power amplifiers can be quite different to that of small-signal amplifiers. Not only are the currents higher, but the effects of temperature and signal level can have a deleterious effect on the bias network. One of the most important concerns to be aware of is that bias points will normally shift in power amplifiers.

5.5.1 Bias changes at the input

The most noticeable impact of a bias point change at the input occurs with a MESFET amplifier, because the gate-source bias is negative and in

small-signal conditions draws no dc current. As a larger RF signal is applied, the instantaneous voltage at the gate can swing above zero volts and the gate-source diode will begin to conduct and draw gate current pulses during that portion of the RF cycle. These pulses correspond to the tips of sine waves, much like the output current waveform in a class-C amplifier. This rectification effect can be intentionally used to self-bias the gate of an FET, in much the same way that grid-leak bias was used in vacuum tubes, although this is rare today except in some oscillators where the input signal swing is constant. The blocking capacitor between the RF input and the FET gate will charge to the value of the peak RF voltage V_1 on positive signal swings as the FET begins to conduct, when the gate is clamped at its turn-on (assumed zero volts). For that portion of the RF cycle when the FET does not conduct at dc, the blocking capacitor will remain charged at the value $-V_1$. The gate voltage is therefore $V_{GS} = V_1(\cos\omega t - 1)$, and the peaks of the signal swing are clamped to the FET turn-on voltage (zero), with average bias voltage of $-V_1$.

When we apply an external gate-bias voltage, we will encounter this problem only if we do not consider the requirement to support a component of the dc current. As the RF swing increases, the conduction sine-wave tips increase in both amplitude and conduction angle as the FET conducts, and an increasing dc current starts to flow. Consider the bias network in Figure 5.58(a), where the bias at node A is initially –2V, set by a resistive divider to a –5-V supply.

This bias could be achieved by setting R1 = 3 kΩ and R2 = 2 kΩ. The amplifier could then be measured and tuned under small-signal conditions. However, when a larger RF signal is applied, as the voltage on the gate

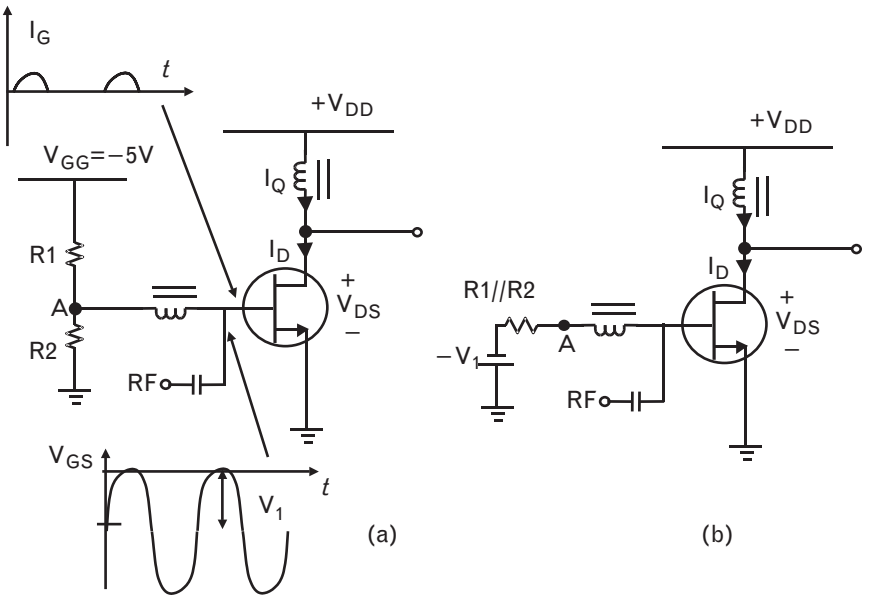


FIGURE 5.58
(a) The gate bias network for an FET. (b) The Thevenin equivalent network.

swings larger than 2V zero-to-peak, the gate voltage becomes positive for portion of the cycle and the gate-source diode turns on. The gate current, previously entirely reactive because of the gate capacitance, now supports a forward conduction component as well, and an average value of dc current must flow through the RF choke. Suppose this is 1 mA in value.

Thevenin's theorem is useful for characterizing the bias network. Thevenin's theorem states that a linear network is equivalent to a constant-voltage source V_0 in series with an impedance V_0/I_0, where V_0 is the open circuit voltage and I_0 the short-circuit current computed at a pair of terminals. In the case of the bias network shown, the Thevenin equivalent is a voltage source of –2V in series with a bias resistance of 1.2 kΩ (R1 in parallel with R2).

As the RF voltage is applied, the 1-mA current generates a dc voltage drop of –1.2V across the bias resistor, lowering the bias voltage at node A and the gate to –3.2V. The unintentional effect of the bias resistance is to drive the FET towards pinch-off as drive is applied.

We will see in a later section that some designers have attempted to utilize this bias shift as a self-regulating mechanism to reduce distortion. In some devices, such as oscillators, the use of self-bias can even remove the necessity for a dual (negative) supply to bias the gate. In most instances, however, a bias change at the input is unwanted since it shifts the gain and the modeling assumptions made in the initial design.

There are a couple of solutions. One solution is to lower the Thevenin-equivalent bias resistance by using lower resistors between the –5-V supply and ground. Using R1 = 300Ω and R2 = 200Ω increases the standby current that flows through them from the rail to ground, but lowers the Thevenin equivalent bias resistor to 120Ω, so that the gate voltage only drops –0.12V from its static point. This is wasteful because of the additional power dissipated in the bias resistors.

Another solution to clamp the gate bias voltage at –2V is to use a Zener diode with a breakdown voltage of –2.1V at the gate. If connected in reverse bias as indicated in Figure 5.59, when the gate voltage falls below –2.1V, the breakdown voltage of the Zener diode is exceeded and the diode goes into avalanche. Then, the diode is a low resistance path to ground and can source the dc gate current required, holding the voltage across it at a very precisely defined value. A cheaper solution would be to stack three diode voltage drops in shunt with the gate, as also shown in the figure. Normally, the diodes are not conducting because the total turn-on voltage of the three diodes is –2.1V; only when the gate voltage falls below that threshold do they start to conduct and provide a low resistance path to ground to source the current.

The same effects occur in a bipolar transistor, but are usually less surprising because they are anticipated in the case of a bipolar. Because the base-emitter diode is always forward biased in active device operation, the base bias network is designed to support quiescent forward current into the

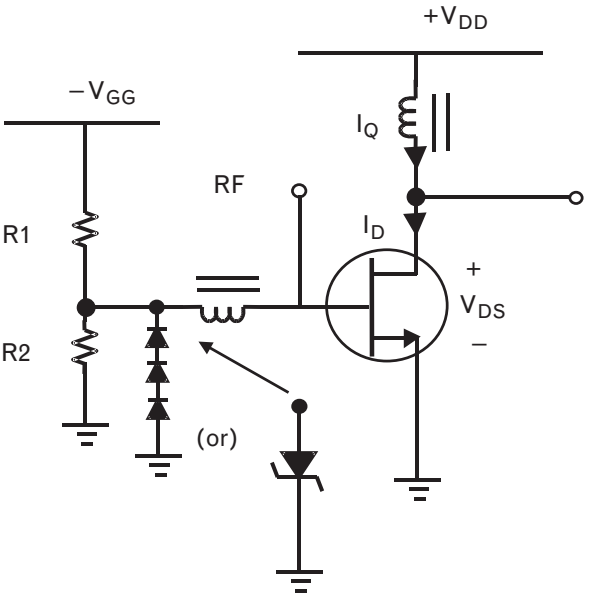


FIGURE 5.59
Solutions to clamp the gate voltage at a fixed bias point.

transistor. However, if the transistor is overdriven, the base current required can exceed the design parameters, and as for the MESFET, the base current that flows through the equivalent network bias resistance generates a voltage drop across it. For bias current that flows into the base (NPN case), this drop tends to reduce the effective voltage across the base-emitter junction.

Similarly, for enhancement mode pHEMTs and JFETs, where the gate is biased positive and some forward conduction into the gate occurs even with no RF drive, care is required to ensure the bias current remains in a safe operating region. As the RF drive level increases, rectification of the gate current is bias circuit dependent. For this type of HEMT, using a *large* Thevenin equivalent series bias resistor will limit the dc gate current to a safe value as it will reduce the positive bias on the gate as the drive increases. Although a large resistor will have a large impact on the gate bias voltage, using such a series resistor has been found to have minimal impact on the RF output power. For the Agilent ATF-54143 device, for instance, a series resistor of 10 kΩ is in fact recommended to ensure the maximum gate current remains limited [16, private communication]. This is counter to the practice in most power (depletion-mode) MESFETs where the resistor will generally be as small as possible to keep the gate voltage constant and limit thermal effects.

Frequently, a small series chip resistor is also included as part of the gate or base bias network to ensure stability, especially at low frequencies. It can also improve the isolation of the bias circuit from the RF at low frequencies. This resistor is particularly important when biasing the gate of power FETs, so that should the FET net input resistance become negative (due to oscillation), the total external resistance will remain positive and prevent

bias-circuit oscillation and unforeseen, excessive, gate current. Plotting the stability circles with the bias ports as inputs, across a broad bandwidth, can be revealing, and can indicate the need to add stabilizing resistors since stability is required near the short-circuit impedance point on the Smith chart around dc.

5.5.2 Bias changes at the output

When power is drawn from an amplifier, it is logical to expect the bias point at the output to shift as well. Figure 5.60 shows the measured drain current for a typical power MESFET at different gate bias points, as the incident RF power is increased at the gate. It is apparent that in this instance, the drain current, and thus related parameters such as gain and the device efficiency, are also functions of the input power. To complicate matters further still, the dc current can either increase or decrease with power, depending on the input bias condition. The shift in bias point will

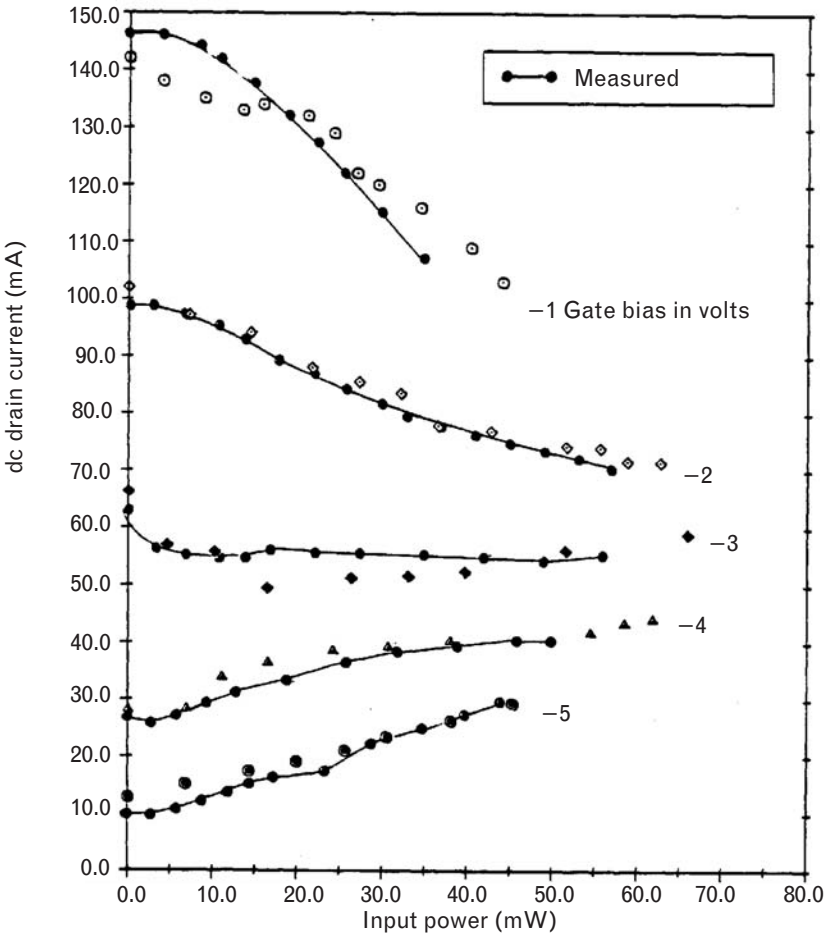


FIGURE 5.60
The dc drain current for a MESFET at different gate bias points, as a function of applied input RF power.

be further exacerbated if a series resistor in either the drain, or quite commonly the source, is used, as this will shift the drain bias voltage and in the case of the latter, the gate-source bias voltage as well.

In many instances, however, this change in bias current, which also occurs for a bipolar device, is an inevitable consequence of amplifier operation. Figure 5.61 shows the transfer characteristic for a MESFET relating the gate voltage to the output current. Although this is drawn with a square law characteristic between voltage and current, resulting in a nonconstant g_m, the principle applies equally for linear devices with constant g_m, as it also would for bipolars. First consider class-B operation, when the device is biased at point "a." As the input power increases, the gate voltage swing becomes larger and drives the device into higher drain currents during the on-cycle. As a result, the average drain current, or its quiescent dc value, must also increase proportionally to the peak value of the half sine-wave, as I_p/π. The opposite occurs when biased at point "b," where during the positive-going gate voltage the device is driven into forward conduction and the resulting saturation of the device and its series resistance limit the resulting peak current swing at the output to the maximum device current. On negative-going gate-voltage excursions, the drain current is able to swing negatively, so that an asymmetrical drain current waveform results, with larger negative peaks than positive peaks, and a current waveform that progressively become flat-topped. Consequently, the average value of the drain current, its quiescent or dc value, falls from its zero-drive value. Finally, when biased at point "c," both the cutoff and saturation effects occur simultaneously and these competing effects tend to

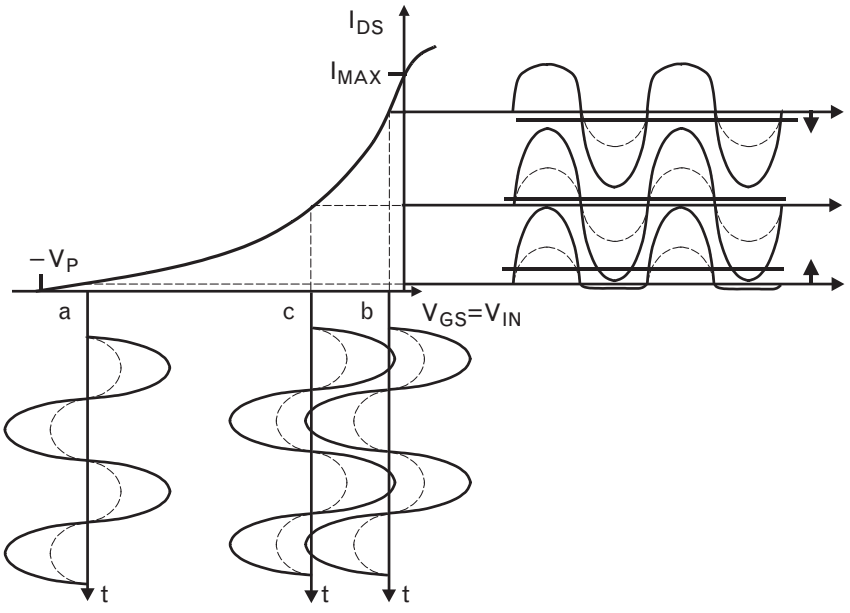


FIGURE 5.61
The relationship between the input and output bias points as the output current changes is illustrated using the transfer relationship for the device. Here, the MESFET transfer characteristic shows the change in dc output current with the magnitude of the input voltage swing at three different bias voltages.

compensate each other. Ultimately, the dc current can remain relatively constant.

The latter illustrates some interesting effects. With nonconstant g_m as shown, the drain current waveform is nonsinusoidal and its average value will shift somewhat as a result of the distortion in the output waveform itself. This will not occur with constant g_m, as then the output current is purely sinusoidal and symmetrical about the bias point. Only during the grossly distorting effects of clipping and saturation will the waveform begin to distort. Even then, however, it is conceivable that the symmetry of the distortion can be maintained, with "equal" clipping and saturation on the negative and positive excursions of the drain current, respectively. Ultimately, the drain current will become square, at which point the even harmonic components will be suppressed relative to the odd harmonic components [from (5.29)]. This *sweet spot* in the output even harmonics can be observed in Figure 5.4 over an interval for which these effects predominate; eventually, other more nonlinear effects predominate and the simple square wave analysis becomes invalid. In gross saturation, when the device is driven between zero and maximum, effects such as the capacitive current from charge stored in the output and feedback capacitances can become substantial. The total drain current can even become negative (i.e., flow out of the device) as a result of these capacitances discharging, even though the current component from the internal *current source* is forced to zero.

As an aside, it is worthwhile noting that in the event that the drain current waveform can be forced to take a square-wave shape, a limiting amplifier will result. The zero-to-peak component of the sinusoidal current waveform just prior to squaring of the current waveform will be $I_{MAX}/2$, while the zero-to-peak component of the square-wave current waveform in limiting operation will be $2I_{MAX}/\pi$. The transition between linear operation and limiting operation is then as sharp as possible, with the difference in fundamental current between the two regions of 27%, or 1 dB. This is clearly a smaller increment than for other bias points, where the transition region is extended over a much greater range of input powers, and compression of the signal swing occurs at one end of the load line before the other. When the onset of clipping and saturation of the current waveform does not occur simultaneously, hard limiting of the output power is much more gradual with increasing input drive.

5.5.3 Bias considerations with power devices

Apart from the shift in bias point that occurs when a device delivers significant output power (compared with its dc input power), there are a number of other factors that need to be considered when a small-signal design is translated to large signal.

5.5.3.1 Bias network to support high currents

High device currents require bias lines capable of supporting high dc currents. Many has been the time when the designer would power on his device and find zero current, only to discover when looking through the microscope that the bias lines acted as fuses rather than bias rails! A 1-mil (25 micron) diameter gold wire fuses at 1 amp. A 250-micron copper microstrip line (68 micron thick) fuses at 4.5A. Likewise, high-impedance, high-inductance microstrip transmission lines are typically very thin—so beware of the dc current they must support. To help avoid high currents due to fault conditions, many power supplies are designed to contain either current limiters, or, in the case of FETs, interlock circuits that prevent the application of any drain voltage until the gate voltage has first been applied. Because of their high transconductance and potential instability at low drain voltages, the drain voltage to power FETs should generally be turned on and off only after the gate has been biased at pinch-off. This prevents the full I_{DSS} bias current flowing through the device, potentially destroying it during turn-on or turn-off.

A further consideration is that although bias networks are typically designed with lowpass characteristics, rarely is attention given to their cutoff frequency. In the case of power amplifiers, where the series bias inductance cannot be as high as could be achieved in the small-signal case, the cutoff frequencies will be even higher. This allows RF currents to leak out the bias ports of the device, and possibly feed back to other parts of the circuit. This is particularly troublesome in nonlinear devices, where mixing or intermodulation products at the difference frequencies circulate within the device and thus potentially out the bias ports. In the case of intermodulation distortion, this leakage of the difference frequency from the output bias back to the input port can cause modulation of the device at the difference frequency, and even causes asymmetry in the intermodulation products themselves. Such secondary modulation at the difference frequency can affect both the amplitude and phase of the main signal and generate mixing sidebands that add and subtract from the primary intermodulation distortion generated within the device itself.

Such effects are known as memory effects, and they are caused by the impedance of the system at the envelope (or modulation) frequency. In two-tone measurements, they can result from a significant reactive component in the terminating impedance at the difference frequency [17]. Envelope impedances will have long time constants, and they are not only associated with the bias network, but can also be thermal related. Some simple principles should be used to prevent such memory effects. One general principle is to decouple RF and dc grounds. The use of voltage regulators and op-amps to provide a good short-circuit to the envelope components that leak out the bias ports can prevent any regeneration that will cause degradation of the *adjacent channel power ratio* (ACPR) in radio systems. So too

will the replacement of narrowband RF short-circuits created using quarter-wave stubs with capacitive termination to a real ground. Better RF and dc decoupling is achieved using the layout shown in the right of Figure 5.62. Capacitor C_1 needs to have a high self-resonant frequency and needs to be small enough so that it does not become inductive at the desired frequency. The large bypass capacitor C_2 for the supply should be supplemented with a small capacitor in parallel to avoid the self-resonance of the large capacitor causing a high reactive impedance in series to the ground.

5.5.3.2 Temperature effects on bias design

In the case of the bipolar transistor, there are two effects that can shift operation of the device from class-C at low temperatures to class-A at high temperatures.

The first is the reverse leakage current through the reverse-biased collector-base junction diode, I_{CBO}. In a silicon bipolar transistor, for instance, the thermally generated component of this current doubles for each 8°C to 10°C increase in operating temperature.

The second effect is that the base-emitter turn-on voltage of the transistor decreases with rising temperature. This has the same effect on the collector current as increasing the dc base voltage by 2 mV per degree Celsius. If the base-emitter voltage were fixed, the collector current would rise exponentially with temperature.

Both of these effects can be minimized through the bias stabilization techniques discussed in Section 1.8, and in particular by reducing the Thevenin equivalent input resistance of the resistive divider on the base and increasing the emitter resistance.

5.5.3.3 Use of bias to control output power

In digital cellular systems, the handset transmitter power is adjusted depending on the received signal strength at the base station. The

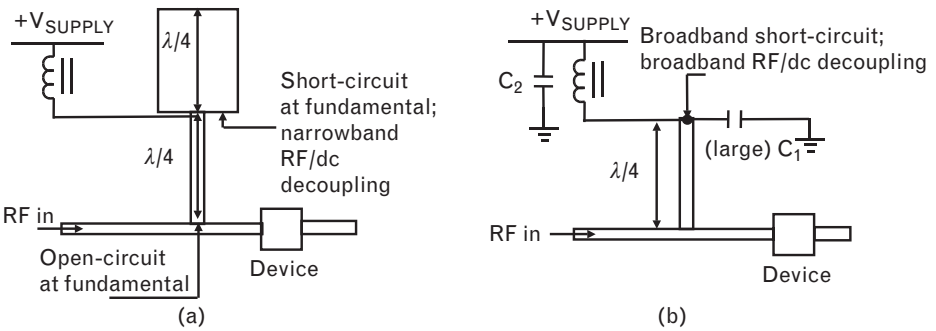


FIGURE 5.62 *(a) An open-circuit quarter-wave stub used to achieve an RF short-circuit that can be used as a bias feed point. This is a narrowband bias network since the RF short-circuit only exists around the fundamental frequency. (b) A better broadband decoupling of RF and dc is achieved using a real dc ground.*

requirement on transmit range in GSM systems is 28 dB, and in wideband CDMA, where it is important that each signal appears to adjacent signals as "noise" the requirement is 71 dB (from –50 dBm to at least +21 dBm).

Systems with a modulation envelope that varies in amplitude require very linear amplifiers. These are typically class-A to avoid spectral regrowth and to meet the adjacent channel power requirements imposed. However, because of the very poor efficiency of the class-A amplifier at low drive levels, the bias is sometimes dynamically adjusted to sense the output power and to reduce the dc power when high bias levels are not required to support the RF power requirement. *Digital signal processors* (DSPs) can be used with an envelope detector for shifting the bias point depending on the instantaneous power requirement [18]. Although transmitter linearity is less important in GSM where the GMSK modulation scheme used has a near-constant envelope and spectral regrowth is minimal, the use of bias for power control and to maintain mobile talk time is always essential.

Although shifting the bias point changes the optimum load impedance, when the power is being reduced from the optimum, the chief concern is either to maintain linearity or to increase efficiency. This can be achieved by adjusting the bias, although as the knee of the I-V curve comes closer to the bias point, care is required to avoid voltage-limited operation and reduce the signal swing accordingly.

5.6 Distortion reduction

In Volume I, Chapter 3, we see that distortion directly impacts the dynamic range of a system because it determines the maximum signal levels that can be handled. In particular, intermodulation distortion is particularly undesirable because it falls in-band and typically in adjacent channels. Third-order intermodulation distortion results from the third-order nonlinearity of the device transconductance, and also from remixing of the second harmonic and the fundamental signals within the device.

Third-order intermodulation distortion (IMD or IMD3) can be measured directly using two tones at the input of an amplifier. Two equal level tones at frequencies f_1 and f_2, spaced a small difference frequency apart, are simultaneously input into the device. It is important that the two sources used to generate the tones contain low distortion products themselves (low source IMD), since that will not only contribute to the output IMD but will also vary as the input power is adjusted. The two sources must also be well isolated from each other, for instance, with separate ac power lines and circulators in the RF path, to ensure no IMD results from secondary sources.

In theory, the two tones will be amplified equally and appear at the output, together with equal third-order distortion sidebands at $2f_2 - f_1$ and $2f_1 - f_2$. The power of these tones is compared to the power in the

fundamental tones at the output, and the result in dBc (decibel relative to the carrier) measures the third-order intermodulation distortion ratio IMR. This result is specified for the given input power level. For example, in most modern mobile systems, the desired signal must be detected even when an undesired signal from a nearby interfering source up to 60 dB stronger is present. If the undesired signal is to be reduced prior to detection to more than 15 dB below the desired signal, its output intermodulation distortion must be –75 dBc below its fundamental (interfering signal). The assumption is that the interfering signal itself will be rejected by the selectivity (filtering) of the receiver, assuming the interferer is not at the tuned (desired) frequency; however, its third-order distortion may well be at the desired frequency and its magnitude is totally dependent on the linearity of the components.

A spectrum analyzer is the simplest way to measure the per-tone power in distortion sidebands. If each carrier has a sinusoidal voltage amplitude (zero-to-peak) of 1V in a 1-Ω system, the power per tone is $V^2/2R$ or 0.5W. Sometimes a true rms power meter is used instead, and the total value in the two tones is measured. This gives the sum of the two powers of a single tone (i.e., 1W). The *peak envelope power* (PEP) is also sometimes used as a reference. With two equal tones of slightly different frequencies, their envelope varies at the beat frequency, between zero (when the phases of the two tones subtract) up to twice the carrier amplitude (when the phases add). Since power is proportional to voltage squared, the PEP in the two-tone case is four times the power in a single tone (i.e., 2W). Thus, if a peak reading power meter is used to specify the input power level at which a given level of distortion occurs, a fictitious 6-dB performance improvement results. The two-tone test is thus quite stringent, since it drives the instantaneous envelope power between zero to four-times higher than the power in a single tone. Use of two unequal input tones can sometimes more realistically simulate an amplifier's response to various modulation formats, since the variation in envelope power will not be as great and the output response can better match the modulation intensity used [19].

Even when the two input tones are equal in level, the two output tones are sometimes unequal. One reason for this was given above in Section 5.5.3.1, where the beat frequency leaks into the bias supply and modulates the device, causing AM/PM conversion. This results in a negative and positive output sideband that adds differentially to the output and cause unequal-level components. The effect of low frequency load impedance on third-order intermodulation distortion is generally poorly characterized.

Another measure of nonlinearity is the *adjacent channel power ratio* introduced into a modulated system by a nonlinear device. ACPR is defined in Volume I, Section 3.2.4.2, as the ratio of the power in a given portion of the signal spectrum resulting from sidebands and distortion to the power in the central carrier. For example, a WCDMA transmitter, which uses

QPSK modulation of the subcarrier, requires better than –33 dBc ACPR at 5 MHz offset and –43-dBc ACPR at a 10-MHz offset, measured in a 3.84-MHz bandwidth.

For a given modulation scheme, a relationship between the third-order intermodulation products and the ACPR at a given power level can be derived. If the measurement bandwidths are the same, then the ACPR can also be calculated from a multitone test. This more closely simulates the amplifier when operating under similar loading to a real communications system. If it is excited by n equal level tones, then the ACPR is the power ratio between the total (integrated) power created by the distortion tones in an adjacent channel to that from the in-channel tones [20]:

$$\begin{aligned} ACPR &= IMR_{2-tone} + 10\log\left(\frac{n^3}{16X + 4Y}\right) \\ X &= \frac{2n^3 - 3n^2 - 2n}{24} + \frac{\mathrm{mod}(n/2)}{8} \\ Y &= \frac{n^2 - \mathrm{mod}(n/2)}{4} \end{aligned} \tag{5.47}$$

In the above, both $ACPR$ and IMR, the ratio of two-tone intermodulation distortion to signal carrier, are in dBc, and mod(n/2) is 0 for n even and 1 for n odd.

5.6.1 The importance of amplifier linearity

For the past two decades, and more intensively over the past few years, there have been numerous studies into techniques to reduce nonlinear distortion. Initially, this was motivated by the desire to use solid-state linear power amplifiers in satellite communications. There, the use of a separate high-power carrier signal to transmit each channel (i.e., SCPC, or single channel per carrier) placed burdensome requirements on transmit amplifier linearity in the (shared) satellite to avoid adjacent channel interference.

Linearity was not considered as critical in first generation or some second generation mobile communications systems because the mobile handsets are spread out within a cell, even though the AMPS system is also a SCPC system. Another reason that both AMPS and GSM can employ transmitter amplifiers operating near saturation is that both the AMPS system, which uses FM, and the GSM system, which uses GMSK, have near-constant modulation envelopes. Schemes with constant modulation envelope are generally types of *frequency-shift keyed* (FSK) modulation. However, many second generation and most third generation systems require the use of both amplitude and phase modulation to efficiently utilize the available spectrum and to obtain high data rates. They

include *quadrature phase-shift keying* (QPSK) and its derivatives, and *quadrature amplitude modulation* (QAM). These have a high peak-to-average ratio, or crest factor, and require the power amplifier to operate linearly to preserve the variation in the carrier. In addition, any scheme that transmits multiple carriers will have variable peaks in the envelope as the carriers add and subtract from each other. This occurs in the base stations of most systems, and in multicarrier CDMA and *orthogonal frequency division multiplex* (OFDM) modulation techniques. As a result, any system nonlinearity causes spectral regrowth and high ACPR. Figure 5.63 shows this with the output power spectrum from a multicarrier QPSK system, measured at low and high input power levels [21, 22].

Consider again the simple two-tone analysis of two signals each of 1V in a 1-Ω system. The total average power in the two tones is 1W, while the peak power is 2W, yielding a peak-to-average power ratio of 3 dB. If the signal is passed through an amplifier with an input 1-dB compression point of 1W, the peaks of the signal will suffer substantial saturation over some portion of the modulation cycle, as they swing up to 2V (corresponding to 2W) rather than the 1.414V corresponding to their average 1-W power. Even though the signal is close to zero volts during the other portion of the modulation envelope, simple arithmetic averaging of the nonlinearity does not work as the compression has already occurred. To keep the system linear, the peak envelope power (rather than the average power) cannot exceed the 1-dB compression point. In our example, this corresponds to reducing the input signal level to operate at an average input power level 3 dB below the 1-dB compression point, known as 3-dB input power

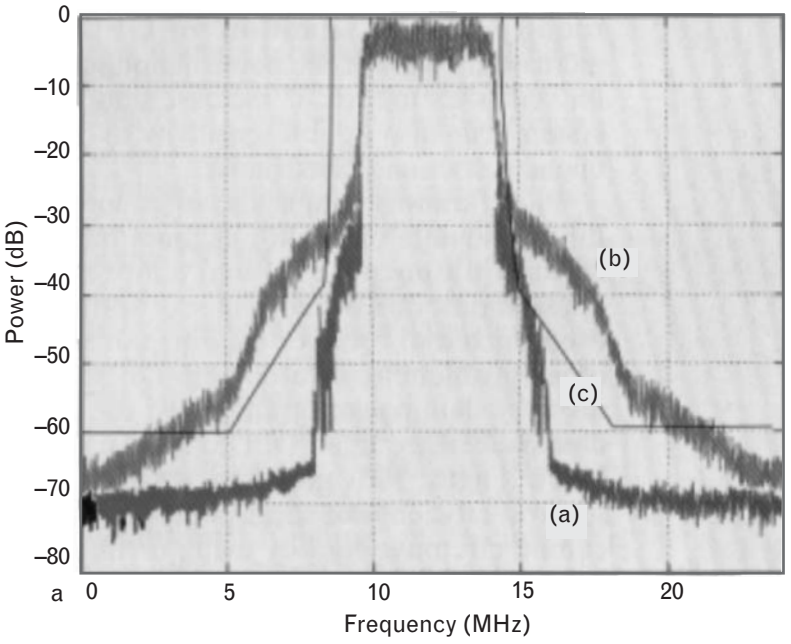


FIGURE 5.63
The output power spectrum of a QPSK256-point OFDM signal from a power amplifier (a) at small-signal, (b) at 10-dB power backoff, and (c) the specified spectral mask. (From: [22]. © 2002 IEEE. Used with permission.)

backoff. Thus, an amplifier designed to operate with 1-W power levels in fact is only operated with 500-mW average power. This can cause problems in class-AB and class-B amplifiers, which may then barely switch on, let alone class-A amplifiers, which will have poor efficiency. Unfortunately, as the number of carriers increases, the peak-to-average ratio becomes even higher and requires correspondingly greater input power backoff. This holds up to the Gaussian limit of 9 dB for an infinite number of carriers whose phases add randomly [6].

To obtain good linearity in an amplifier generally requires low input power. This is acceptable in the receiver amplifier but not in the transmitter, where good efficiency is a major design goal. High power-added efficiency requires either high quiescent current levels or the use of switching-mode amplifiers. There is, therefore, a compromise needed, and numerous distortion correction schemes have been proposed to achieve linearity at operating power levels still large enough to achieve good output power and high power-added efficiencies. Such schemes can be classified as providing either feedforward (or additive) correction, or feedback (or multiplicative) correction [23]. In the former technique, the signal is corrected at the output of the power amplifier by adding in to the main signal an opposing distortion signal. In the latter, the main signal is predistorted prior to amplification so that after the nonlinear process, the compensated signal appears pure. In all techniques, the objective is to be able to operate the amplifier at a higher output power and achieve the same ratio of fundamental signal (carrier) to intermodulation power, or C/I ratio. With TWT amplifiers, up to 6-dB increase in output powers can be obtained for the same C/I ratio, whereas for more linear MESFET class-A amplifiers, 2- to 3-dB improvement can be achieved. Equally important, the dc to RF efficiencies of such amplifiers can sometimes almost double because of operation at a higher output power level.

Because there are a number of recent texts devoted to the entire topic of linearization [6, 19], we will only provide an overview of the basic techniques here.

5.6.2 Operating the amplifier backed off

Figure 5.64 shows that the third-order intermodulation distortion products rise at a theoretical rate of 3 dB for every 1-dB increase in input power. Operating the amplifier *backed-off* simply refers to either lowering the input power to reduce the distortion levels—the principle behind *automatic gain control* (AGC) systems—or to using a larger (higher power) device than necessary to handle the system input power levels. Oversizing the device for the given application will increase the input third-order intercept point of the device and shift the distortion output power curve to the right along the input power axis in Figure 5.64. As a result, the distortion output for

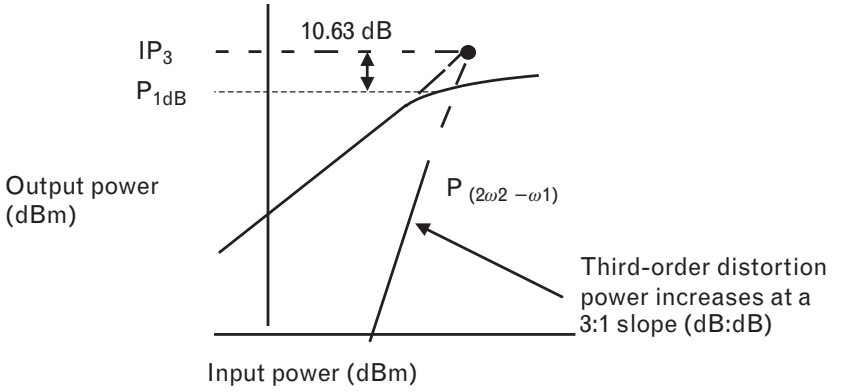


FIGURE 5.64
The relationship between third-order intermodulation distortion, intercept point IP_3, and fundamental frequency input power.

any given input power level is reduced. Such a technique, although widely used in the 1980s for satellite applications, is now unnecessary because of newer techniques, and it was very expensive because of the additional cost of using a higher power device than necessary, and the associated overhead of its power supply, weight, and size.

In the satellite system for instance, if the intermodulation products are not to significantly degrade the noise floor, then they should fall at least 15 dB below it. If an FM sensitivity of 10 dB is required (SNR) and we allow a 5-dB fading margin, the signal should be 15 dB above the noise floor for reasonable quality detection. A *carrier-to-intermodulation* (C/I) ratio of 30 dB is therefore needed for detection. Using the 10-dB rule of thumb relating the 1-dB compression point to the third-order intercept point (Volume I, Section 3.2.4.1), this implies the output power must be backed off by 5 dB (from the 1-dB compression point, where the C/I is 20 dB). In digital cellular systems, the C/I requirement is much stricter, typically between 40 and 60 dB, implying even greater power backoff.

5.6.3 Predistortion

Predistortion attempts to modify the signal before it is amplified, applying to the signal an inverse characteristic to the amplifier itself. Both the amplitude and phase characteristics of the power amplifier should be compensated in the predistorter. Predistortion can be applied at RF, IF, or even at baseband using DSP to adaptively shape the symbols, although our focus here will be on RF predistortion.

We see in Volume I, Chapter 3, that both gain compression and third-order intermodulation distortion are caused initially by the cubic term in the transconductance or system nonlinearity. Thus, perhaps without knowing it, most predistorters attempt to fashion a cubic response that can be added in to the input signal to compensate. Numerous circuits have been published, but often with fairly minor improvement in distortion levels, taking advantage of a sweet spot that may exist over a limited dynamic

range or bandwidth. In their defense, such circuits are generally low cost, low power, easy to implement, and stable, so indeed have their place in handset radios where only minor improvement may be necessary. They can also be a useful adjunct to more complicated system-level correction techniques such as feedforward, where they can reduce the size, for instance, of the error amplifier.

The concept of a simple amplitude predistorter is illustrated in Figure 5.65. An input signal is split into two parallel paths, one that is linear and a second that is nonlinear. A small-signal amplifier in the linear path provides linear gain up to a compression point that is well beyond the compression point of an associated limiting amplifier in the second path. The signal from the limiting amplifier is subtracted from the linear amplifier in an output balun, so the resulting characteristic has a region of gain expansion that can compensate for the gain roll-off of the following power amplifier. Such a scheme compensates for amplitude nonlinearity and could, in principle, also be used to compensate phase nonlinearities as well. It indicates that the best a predistorter can achieve is to correct for nonlinearities over a range of input power levels below the compression point of the main amplifier.

Practical applications of predistortion vary in complexity. One relatively simple implementation uses a forward-biased diode in shunt with the signal, as reported in [24]. Figure 5.66(a) shows the circuit, in which the forward-biased diode provides a shunt resistance and shunt capacitance to the main signal. The R-C combination is proportional to the bias voltage on the diode and the amplitude of the signal swing. The application of a forward bias current to the diode prevents rectification and further distortion of the input signal. This forward bias voltage, in conjunction with the asymmetry in the signal swing of the current through the diode as the input RF voltage swing across it increases, causes an increase in the average dc current through the diode from I_{ds} to I_{dl}. This increases the dc voltage drop across the bias resistor and lowers the diode operating voltage point from S to L in Figure 5.66(b). The bias point shifts along the dc load line as determined by the value of the bias resistor. As a result, and as given by the slope of the diode I-V curve at the new operating point, the incremental shunt resistance of the diode increases with the voltage swing. This reduces the

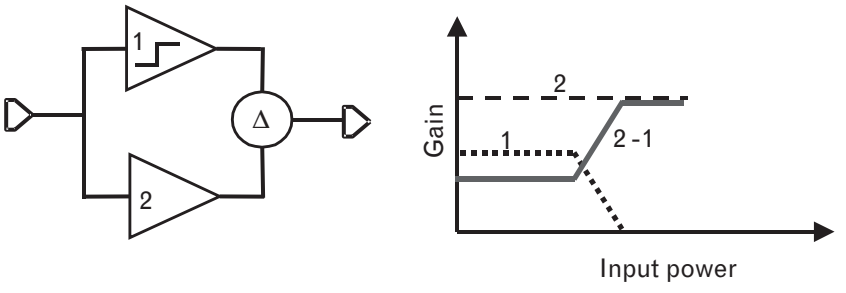


FIGURE 5.65
An example of an RF-type predistortion limiter and its gain characteristics.

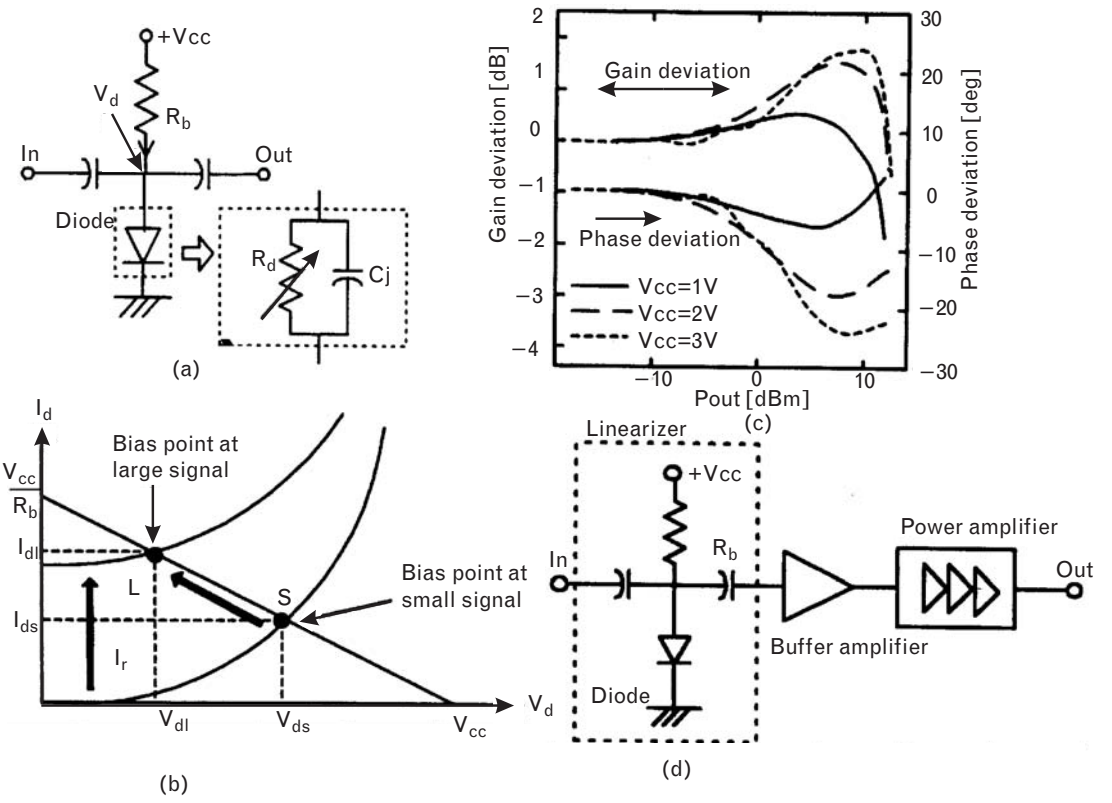


FIGURE 5.66 *(a) A predistorter linearizer. (b) The shift in diode bias point with applied RF input power. (c) The resultant gain and phase deviation with input power. (d) Block diagram of a power amplifier with the linearizer. (From: [24]. © 1997 IEEE. Used with permission.)*

shunt loss at larger signal levels and provides a positive gain deviation and negative phase deviation with input power, as shown in Figure 5.66(c). The predistorter, therefore, when used as in Figure 5.66(d), is able to compensate for the gain compression of the power amplifier and any positive phase slope as the input power increases, at least over a limited range. Improvements of 5 dB in the *adjacent channel power* (ACP) of a $\pi/4$-QPSK signal have been reported at the 1-dB compression point. A similar circuit using a shunt FET rather than a diode is suggested in [25]. This uses a biased FET connected in shunt with the gate of the main amplifier FET to null out its third-order distortion.

A slightly different variant of this circuit uses two shunt diodes connected in shunt as an antiparallel pair. In theory, the even-order distortion circulates completely within the pair and only the odd-order distortion remains. If this distortion component is reinjected into the main signal path prior to the power amplifier, then similar improvements can be obtained to the above.

Two other types of diode predistorter circuit are shown in Figure 5.67. In the first, the simple back-to-back diode pair shunted by a resistor is inserted in series with the signal. As the input power increases, the series resistance of the diode decreases as its operating point is moved up the diode I-V curve. In shunt with the diode capacitance, its insertion loss becomes smaller and the phase decreases. The second circuit is inserted in shunt with the signal. When either is inserted prior to an amplifier with opposite characteristics, the cascaded AM-AM and AM-PM curves can be flattened and linearity improved over a range of input power levels, generally just below the 1-dB compression point of the amplifier. The size of the diode can be scaled in order to better match the desired characteristics. Although the degree of flattening is generally insufficient to totally compensate, improvements of around 5 dB in the ACPR have also been reported [26]. With modulation formats with large peak-to-average ratios such as n-QAM (i.e., with highly varying modulation envelopes), the improvement is less because the dynamic range of the amplitude fluctuation exceeds the linearization range of the predistorter. Like all diode circuits, the effect is also temperature dependent.

A second type of predistorter, known as a self-phase distortion compensator, is reported in [27] and shown in Figure 5.68. Although not strictly a predistorter in that the inverse compensation is provided *after* the power amplifier, the principle is similar in that the in-line signal is modified in a multiplicative way by a phase characteristic inverse to that of the main amplifier. In Figure 5.68(a), the phase deviation of a *common-source FET* (CSF) is positive with input power (i.e., as the incident power increases, the phase of the output signal increases even though the input phase remains constant). This is principally the result of an increase in the output conductance of the FET as the signal increases. However, for a *common-gate FET* (CGF), the phase deviation is in the opposite sense (negative), because the conductance now lies between the input and output terminals of the FET and there is no longer a voltage inversion between the input and output as there was for the common-source FET. As a result, the common-gate FET can act as a phase distortion linearizer for the common-source FET. There is a difference in power levels at which the cancellation effect occurs in the two devices, but it is somewhat offset because the common-source FET will amplify the signal close to the level required for the common-gate FET. The bias levels at the gates can also be adjusted to provide appropriate saturation levels for the two devices. The resulting system,

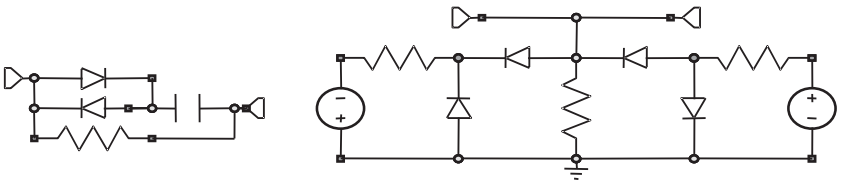


FIGURE 5.67
Two typical diode predistorter circuits.

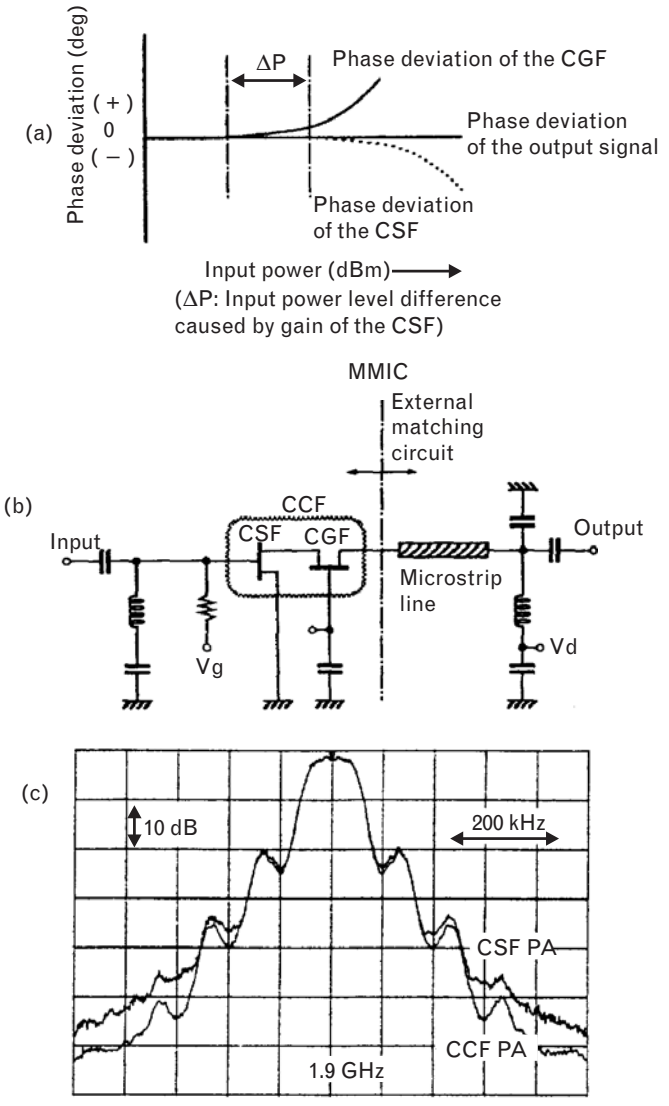


FIGURE 5.68
A pair of FETs used for self-phase distortion compensation. (a) The principle of combining a common-source FET with a common-gate FET to achieve a reduction in phase distortion. (b) Application in an amplifier circuit. (c) Resultant output power spectra of a common source FET with and without the predistorter. (From: [27]. © 1995 IEEE. Used with permission.)

shown in Figure 5.68(b), can be achieved with a dual-gate FET, which intrinsically is just a cascode connection of two FETs (CCF). The gate of the second FET, the common-gate amplifier, is short-circuited at RF in order to minimize any phase shift in that amplifier.

This technique adjusts for phase distortion only, and not for amplitude compression. However, AM-PM conversion near saturation is a problem in a number of phase-modulated systems, and its reduction alone can help improve the adjacent channel interference of such systems. Figure 5.68(c) shows the improvement achieved can be up to 10 dB in alternate channels.

A related use of the dual-gate FET as a true predistorter is reported in [28]. Here, the bottom FET is held at a constant bias and the signal applied to the top FET. With increasing input power, up to 5 dB of gain expansion

and 15° of phase change could be introduced into the signal path. With the bottom FET biased at near-zero drain voltage, its drain-source resistance increases with input power while the transconductance of the top FET increases, providing gain expansion due to a class-B effect when the device switches on. By tailoring the relative size of the two devices and adjusting the bias of the top FET, the dynamic range over which correction is applied could be adjusted to over 30 dB. Up to 10-dB improvement in ACPR for a CDMA transmitter signal was achieved.

A third predistortion technique is known as interstage second harmonic enhancement [29], and is a system technique that predistorts the signal between two cascaded amplifier stages. In this case, the interstage network is designed to adjust the second-harmonic component of the signal produced by the driver amplifier, so that it remixes with the fundamental in the following power amplifier to produce its own third-order distortion. There, the directly generated third-order distortion (from the g_{m3} components of the transconductance of the driver and power amplifiers) is cancelled by this second-order effect (i.e., the direct mixing of the second-harmonic and the fundamental to produce a signal at $2f_2 - f_1$). The second-harmonic needs to be adjusted in amplitude and phase by the interstage network to ensure this cancellation of the directly generated product can occur. We should note that the second-harmonic component is generated by the second-order nonlinearity of the driver amplifier, and the mixing in the power amplifier also arises predominantly from its second-order nonlinearity. The third-order distortion arises predominantly from the third-order nonlinearities of both stages. Figure 5.69(a) shows the experimental setup, and Figure 5.69(b) shows the resulting 15-dB reduction in spectral regrowth for a $\pi/4$-QPSK signal at 1.8 GHz. A modification of this technique is to inject the second-harmonic into the power amplifier from a parallel predistorter circuit that samples the input signal. In this case, the predistorter can be realized using a class-B amplifier to generate the appropriate level of second-harmonic distortion.

5.6.4 Feedforward cancellation

Feedforward cancellation of distortion has the advantage that all spurious signals over a wide range of input power can be cancelled. It also has less of the stability problems that are inherent to some feedback systems, and it can achieve cancellation over broad bandwidths if broadband adders and couplers are used. Nor is the improved linearity achieved by reducing gain, as with a feedback system. However, it is a complicated, system-level technique that requires integration of a number of components as it requires an extra error amplifier and appropriate phase and gain compensation.

Figure 5.70 shows the principle. The system subtracts a sample of the output signal from a pure sample of the input signal to leave only the distortion component. The output of the main amplifier, in the bottom arm

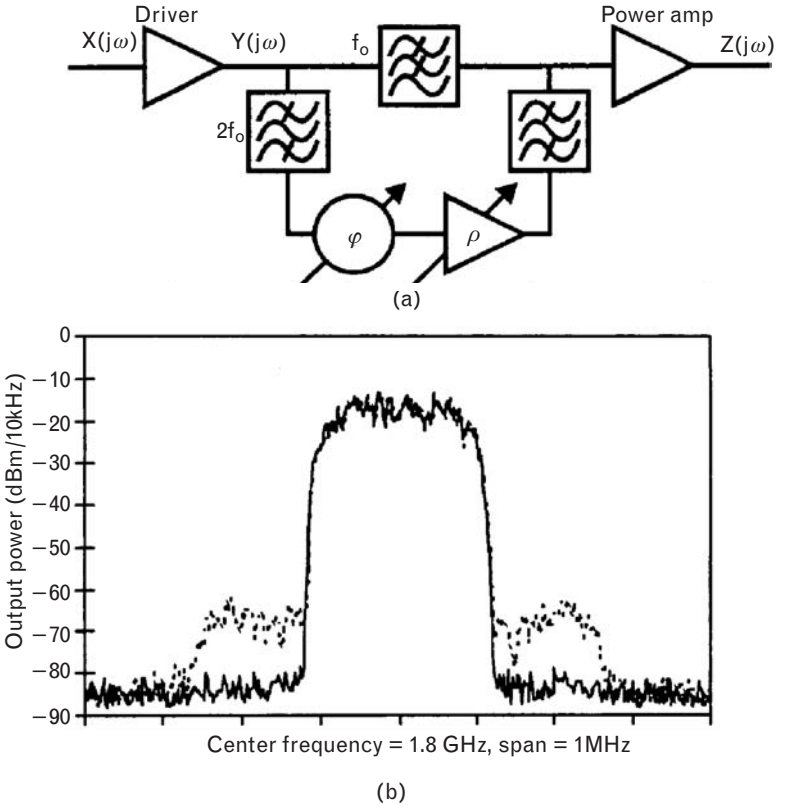


FIGURE 5.69
(a) Block diagram showing the principle of interstage second-harmonic enhancement to reduce distortion. (b) Resulting output power spectra. (From: [29]. © 1998 IEEE. Used with permission.)

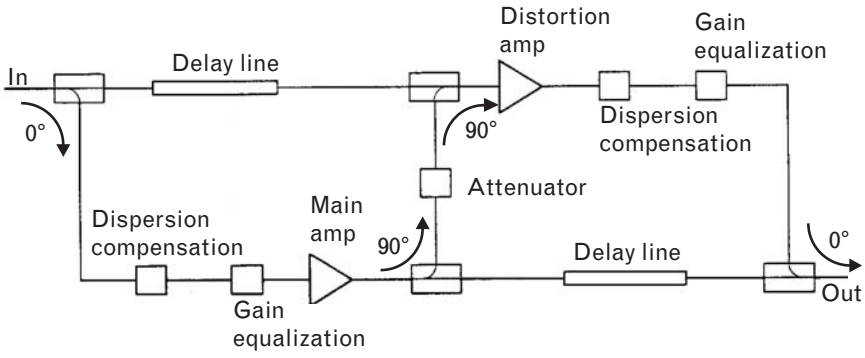


FIGURE 5.70
Principle of feedforward cancellation to reduce amplifier distortion.

of Figure 5.70, contains both the required, high power signal and associated distortion. By subtracting an attenuated sample of this signal from the original signal, in the top arm of Figure 5.70, only the distortion component remains. This signal is then phase and gain equalized so that it in turn can be subtracted from the output of the power amplifier, leaving only the amplified main signal at the output.

The system in Figure 5.70 includes delay lines to compensate one signal for the time delay in RF processing of the other. The RF processing includes dispersion compensation (to adjust the phase of the signal for delay mismatch in the various transmission lines) and gain equalization (to

compensate for coupler losses and amplifier ripple). Both 0° and 90° couplers are used to split the signal and recombine it with the correct phase sense (addition or subtraction). Improvements of up to 7 dB in third-order intercept, and 20 dB in the amplitude of the third-order products have been reported.

However, feedforward cancellation is rarely ideal. Cancellation is limited at higher power levels (nearing compression of the main amplifier) by the size of the secondary amplifier, whole size, in turn, is determined by the coupling that can be used at the output. This amplifier typically requires compression characteristics not much lower than the power amplifier itself, particularly when the insertion loss of the output coupler on its signal is accounted for at the output. Also, amplitude imbalance occurs as a result of amplifier ripple, mismatch, and coupler roll-off. Phase mismatch occurs as a result of dispersion in the transmission lines and couplers, delay errors between the two arms of the system, and phase offsets introduced by the addition and subtraction within the system. These errors reduce the suppression of distortion as shown in Figure 5.71. Even with no phase error, an amplitude imbalance of 3 dB reduces the maximum achievable suppression to 10 dB below the original distortion levels; phase errors greater than 15° of imbalance have the same effect.

5.6.5 Device modification

Perhaps the most obvious place to look to reduce distortion is within the device itself. In fact, the most successful amplifier designs begin with the selection of a good device.

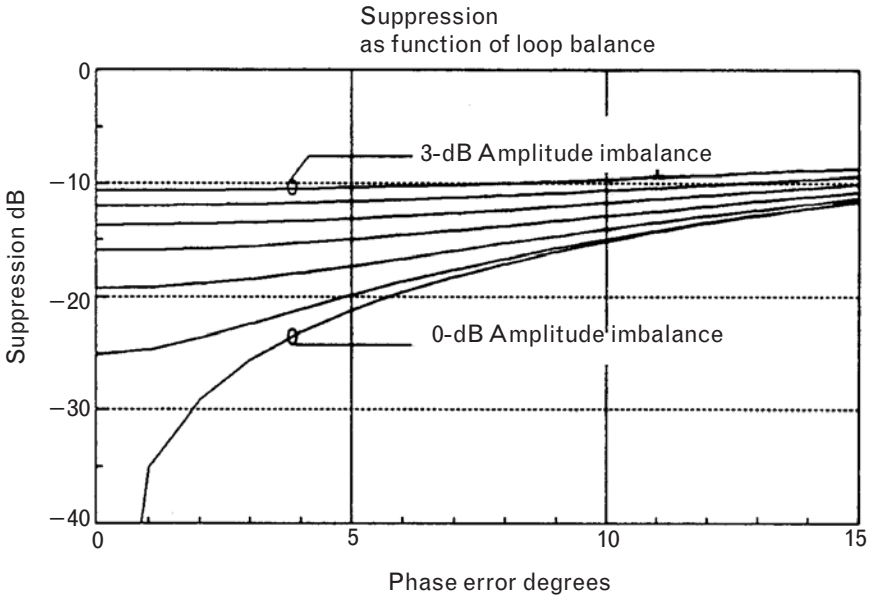


FIGURE 5.71
Suppression of the feedforward cancellation system as a function of loop balance.

The first principle in minimizing distortion is to choose as linear a device as possible. A linear device is one in which the I-V output curves are evenly spaced with gate voltage or base current, so that the device transconductance or current gain is linear. Some MESFETS, and most HEMTs and HBTs, are good choices to start with. Samelis in [30] measured the intermodulation distortion from an HBT, and he and a number of other authors concur that HBTs, in general, have good linearity and low dc power consumption. The reason suggested in [30] is that the base-emitter and base-collector contributions to the distortion current also partially cancel. However, Maas [31] suggested that the output current components generated by the resistance of the base-emitter junction diode partially compensate the components generated by the capacitance of the same junction. More recent work [32] concurs that the exponential nonlinearities are cancelled internally within the device itself and suggests that the residual nonlinearity due to the transconductance can be linearized using a series emitter degeneration resistor. In any event, the choice of a more linear device is the simplest and most effective start to minimize distortion.

Consider, for example, the IBM 43RF0100 SiGe HBT transistor shown in Figure 5.72. With an f_T of 15 GHz measured at 3V, 5 mA, the device can be used in most modern wireless systems. Although this device is not a classical "power" device, it can be used either in a receiver LNA or in the transmitter amplifier of a CDMA system where the power requirements are moderate. Linearity in the receiver is just as important as for the transmitter. In the transmitter, the concern with linearity is to avoid transmitting sidebands that fall in adjacent channels; in the receiver, linearity is important to prevent intermodulation distortion from two, strong unwanted signals generating a third-order product that can swamp a desired weak channel.

Figure 5.72 shows that this device has both low noise figure and good linearity. Its output power capability can be derived from (5.24) since it depends on where the device is biased. Assuming a V_{SAT} of approximately 0.5V, if the quiescent current is 10 mA and the supply voltage 2.5V, the maximum linear output power will be approximately 10 mW or 10 dBm. At 2 GHz, the input third-order intercept point is 10 dBm, so with a typical gain of 12.5 dB, the output third-order intercept point is 22.5 dBm. This is a fairly high intercept point relative to the estimated 1-dB compression point—a difference of 12.5 dB compared with the theoretical 10 dB—so the device is reasonably linear.

Similarly, the Agilent ATF-54143 is an example of an enhancement mode pHEMT that has excellent linearity. When biased with a dc drain current of 60 mA and a drain voltage of 3V, it is capable of a 1-dB compressed output power around 19 dBm, yet its specified output third-order intercept point is 36 dBm at 2 GHz, well above the classical 10-dB difference discussed in Volume I, Section 3.2.4.1.

SiGe high dynamic range low noise transistor

Features

Low noise figure: NFmin ≈ 1.1dB @ 2.0GHz
V_{CE} = 2.0V, I_C = 5mA

Input IIP3 capability: ≈ 10dBm @ 2.0GHz
V_{CE} = 2.5V, I_C = 10mA

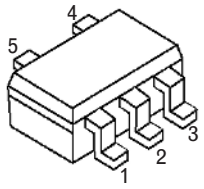
Low operating voltage V_{CE} = 1.0 to 2.5V

Package: SOT353

Description

The IBM43RF0100 is a Silicon-Germanium (SiGe) NPN transistor designed for high performance, low cost applications. Utilizing IBM's SiGe process and packaging technologies, high gain, low noise and exceptional linearity at low power consumption are possible. Assembled in a miniature surface mount package, this product is designed for applications requiring high performance such as LANs, VCOs, and other low noise transistor applications.

Pin diagram



Pin assignments

Pin 1	Base
Pin 2	Ground[1]
Pin 3	Emitter
Pin 4	Collector
Pin 5	Emitter

1. Connection requires a low resistance path to signal ground.

FIGURE 5.72 *Specification of the IBM 43RF0100 SiGe HBT transistor, showing its basic features. (Courtesy IBM Microelectronics.)*

Cancellation effects have also been observed in pseudomorphic HEMTs. These devices do not have constant g_m, but instead exhibit peaks in their transconductance at gate voltages slightly above zero volts. They also have maximum drain currents considerably higher than I_{DSS} when the gate voltage swings positive [33]. When biased class-A, these effects cause gain expansion prior to compression as the input power is increased. In a rather contrary way, the presence of both third- and fifth-order terms in the transconductance (expressed as a function of gate voltage) actually enables cancellation of the third-order products over a particular power range. Output intercept points 15 to 22 dB higher than the 1-dB compressed power are typical, compared with the classical 10 dB expected.

Given that the output current source of a transistor is the dominant nonlinearity, its third-order intercept point is most sensitive to the real part of the output load impedance at the fundamental frequency. The output intercept point tends to be insensitive to its source loading, if there is not too much feedback through the device. Thus, the second principle in minimizing distortion is to choose the correct load line. We have discussed in previous sections the importance of the correct slope and quiescent bias

point to avoid regions of the I-V curve where nonlinear components can be generated, such as clipping (at cutoff), saturation, and nonlinear g_m. These might be termed primary distortion effects. Secondary distortion effects result from harmonic loading of the device. As a general rule, in order to minimize distortion products in the load, all harmonic currents should be terminated in a short circuit at the intrinsic device terminals.

In [30], the fundamental output load of an HBT amplifier is held constant and the second-harmonic load impedance is varied. Figure 5.73 shows the IMD3, relative to the fundamental, as the second-harmonic load is tuned. The distortion varies plus or minus 2 dB as the reflection coefficient is changed, and it is lowest when the reflection coefficient is magnitude one, with angle close to three radians. This corresponds to a short circuit at the collector of the HBT. Although distortion components must flow in the output current of the transistor if the device is even weakly nonlinear, these components can be prevented from generating output power in the load by suitably terminating them at the device. In general, short-circuit terminations for the distortion components are preferred at

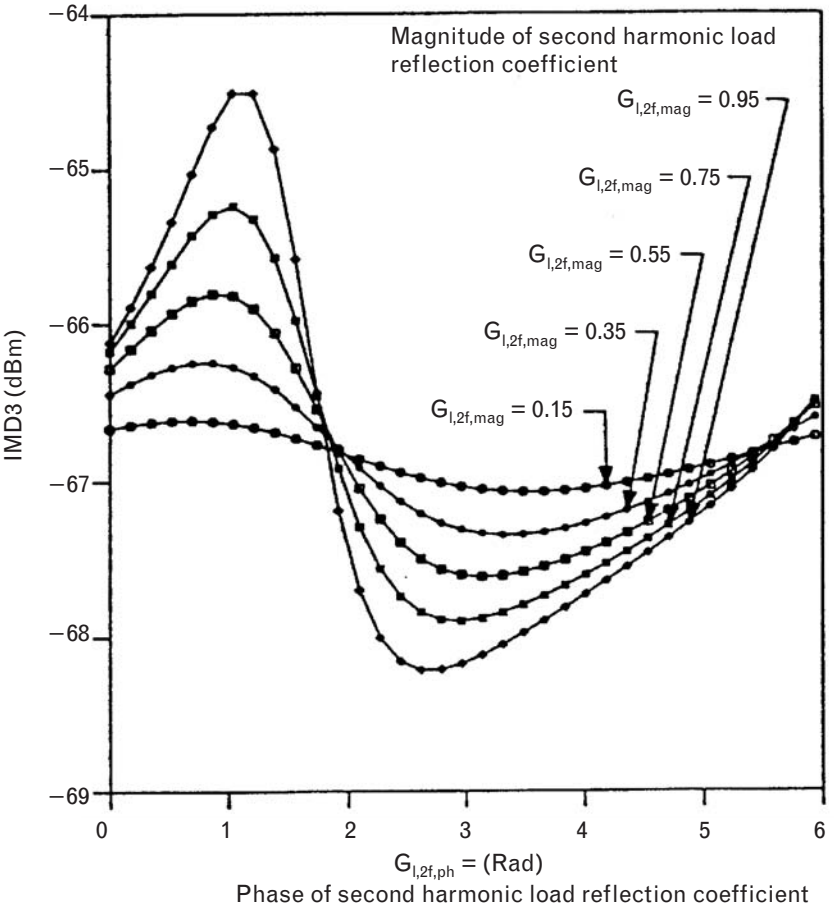


FIGURE 5.73 *Third-order intermodulation distortion measurement of an HBT as a function of its second-harmonic load termination. (From: [30]. © 1992 IEEE. Used with permission.)*

the output current source of the device, in order to ensure zero voltage there and to minimize the effects of any voltage feedback to the transistor input through the collector-base (or drain-gate) capacitance. Such voltage feedback, if incorrectly phased, could remix in the device and generate further nonlinear components through remixing. Of course, if correctly phased, this feedback could be used to reduce distortion. This possibility is called device modification [34] and will be discussed further below.

The third principle in minimizing distortion in the device is to choose an appropriate bias point. Dynamic adjustment of the supply voltage with envelope level is clearly an appropriate technique for minimizing the dc bias power to improve class-A efficiency when signal levels are low, but it should also be considered for impact on linearity. In general, dynamic bias will position the load line in the most linear region of the output I-V curves, where the spacing is constant for constant variation in gate voltage or base current between adjacent curves. Some authors [35] have utilized the change in input bias point with input power level in an attempt to control the intermodulation distortion, but with limited success. Others [36] have intentionally modulated the drain bias voltage of an FET with the low-frequency voltage envelope (difference frequency) caused by the intermodulation at its gate. This feedforward effect reduced the ACPR from a CDMA test signal by 10 dB.

Figure 5.74 shows the effect of base bias resistance R_b on the amplifier gain, fundamental output, and third-order output power as the input power is changed [35]. Because the amplifier is biased class-B, its base current increases as the input signal swing is increased. When a dc voltage source is connected directly to the base ($R_b = 0$), slight gain expansion is observed prior to saturation, with relatively low distortion output. When a

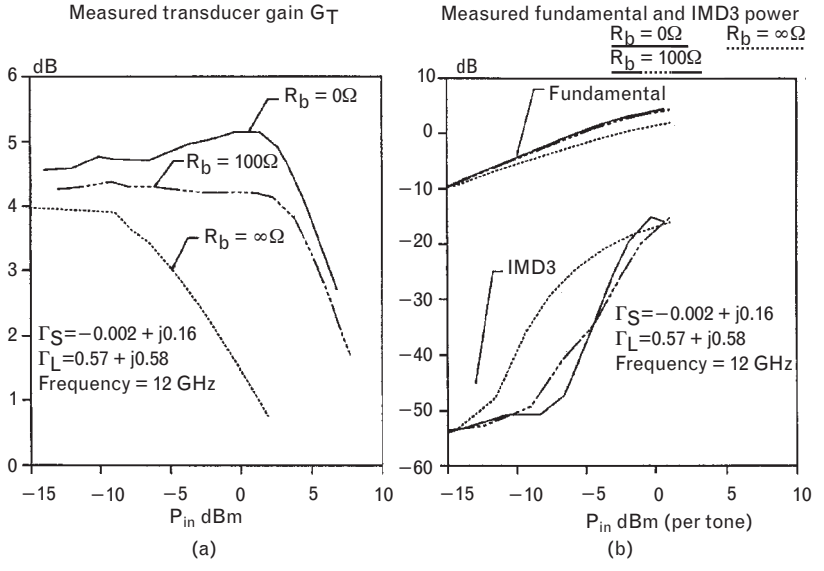


FIGURE 5.74
(a) Measured gain for an HBT power amplifier as a function of incident power with the bias resistor as a parameter. (b) Measured fundamental and third-order intermodulation power IMD3. (From: [35]. © 1992 IEEE. Used with permission.)

resistive divider with equivalent bias resistance of 100Ω is used, the results are similar, although the gain is flatter up to the onset of compression. With a dc current source connected directly to the base, the linearity of the device is considerably worse, as measured by the premature compression of the gain, lower output power, and higher distortion products. These results show that the base bias resistance—and by implication, the shift in input bias point—can impact the performance of the amplifier, not only its gain but also its distortion. Fortunately, the best result is consistent with the sort of bias network we would design for thermal stability of the transistor (i.e., a network with a low equivalent bias resistance). However, the measurements do indicate the importance of being aware of bias point shifts, and they remind us that there may be optimum values of bias resistor that moderate the effect of either the base-emitter—or gate-source—voltage shift that occurs as the base or gate current is rectified as the input power is increased.

Finally, as alluded to earlier with device modification, the judicious use of feedback or feedforward around the device itself can reduce distortion. A use of feedforward is described in [37] and illustrated in Figure 5.75, in which a sample of the input signal is fed forward and applied directly to the drain of a common-source FET. In this instance, a linear amplifier A_2 was used to adjust the amplitude and phase of the fed-forward signal, and an isolator used to apply it to the drain of the FET in the main power amplifier A_1. There, it adds with the output signal produced by the power FET in such a way as to reduce the third-order distortion by several decibels. In one sense, the load line of the device is dynamically modified by the added injected signal so that the load line is "linearized." Baluns or couplers can also be used in place of an isolator to achieve the same injection of the signal into the output of the device. The use of transformers for feedback to modify the device is described in [34] and achieves similar results. The Doherty amplifier [38] also uses a similar principle, although the intent there is to improve the efficiency rather than the linearity. It employs a second auxiliary amplifier to lower the effective output impedance seen by the main amplifier at high power levels. This load pulling effect allows the main amplifier to deliver more current to the load while it is saturated, and thus maintain higher efficiency over an extended power range. The effect of the Doherty amplifier on linearity has not been reported.

As opposed to these lossless feedback or feedforward techniques, resistive feedback applied to the device was considered in Chapter 2 as a way of reducing the gain in order to flatten it across a broad band. Adding negative feedback improves (lowers) the level of third-order products below the main signal by an amount equal to the gain reduction caused by the feedback. In dBc terms, the output distortion is reduced by an amount equal to the gain reduction, for the same fundamental output power. Although a higher input power level is necessary to restore the output signal level to that before feedback, the *relative* output distortion power is reduced. Stated

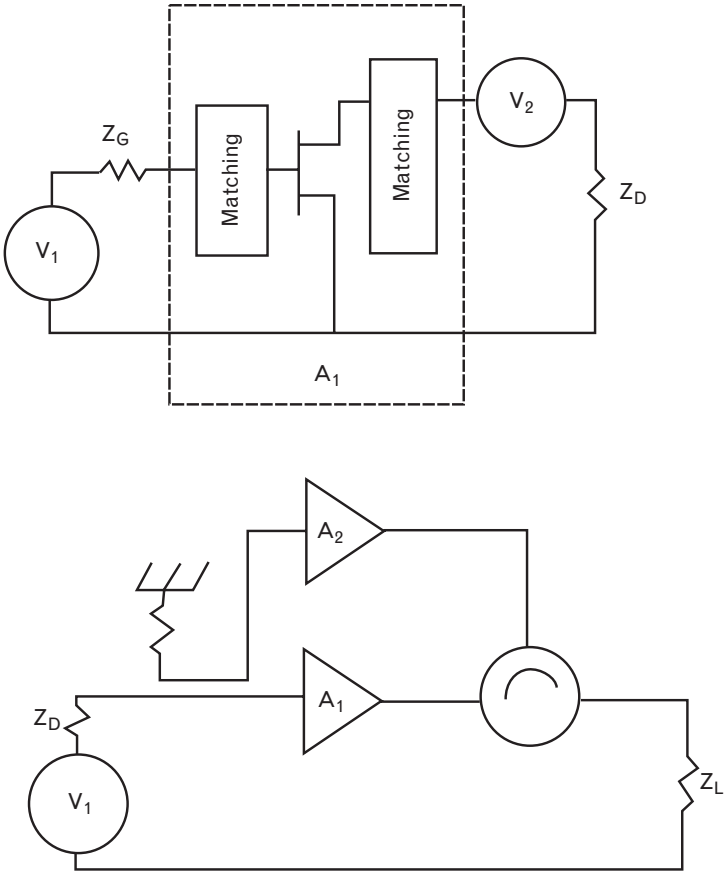


FIGURE 5.75
Principle of modifying a device using feedforward, by applying a suitably phased sample of the input signal to the drain of an FET.

differently, if the gain reduction is *GR* dB, then the output third-order intermodulation products decrease by *GR* dB relative to the fundamental. The output third-order intercept point (OIP3) therefore increases by *GR*/3 dB and the input third-order intercept (IIP3) by 4/3*GR* dB. This assumes the classic 3:1 rise in third-order products with input power. However, in a microwave power amplifier, it is generally impractical to use negative feedback because of losses in the feedback resistors, the limited gain to throw away, and the delay between input and output. Additional stages would also be required to compensate for the reduction in gain and input power. At lower frequencies though, negative feedback can be considered. For instance, adding an emitter degeneration resistor in series with the emitter of a bipolar transistor increases the maximum signal swing that can be tolerated at the base, and thus increases IIP3.

5.6.6 System-level reduction of distortion

Class-B or class-C power amplifiers will probably be required in new mobile systems to maintain the battery life and talk-time that users have

come to expect from the existing second generation mobile systems such as GSM. Because the new systems use modulation formats with varying envelope, linearization techniques applied to the entire transmitter will most likely be required, possibly in combination with some of the ideas discussed above. System techniques include Cartesian loop, polar loop, adaptive baseband predistortion, envelope elimination and restoration, and linear amplification using nonlinear components (LINC). Space permits no more than a cursory overview of a sample of two of these, and for more detail the interested reader is referred to [6, 19, 39]. System-level processing, with the right amount of integration, can frequently yield the best results and the most power-efficient solution, as most of the complexity is built into the baseband processing elements rather than into analog control circuitry.

The Cartesian Loop Architecture [19, 40] is a technique that linearizes an entire transmitter by applying feedback around the entire system. In essence, it detects the transmitted modulation envelope, compares it with the original, and makes an adjustment by forming an error signal. Both the phase and amplitude of the envelope can be preserved by adjusting the amplitude of the I and Q baseband components. The principle is illustrated in Figure 5.76. This system combines upconversion and power amplification so that the whole design is subject to the distortion improvement of the baseband linearizing feedback. A sample of the output signal of the transmitter is downconverted to I and Q signals, which are then used to form feedback error signals by subtracting out part of the original I and Q signals. The error signals are used in the quadrature modulator to predistort the transmitter input to keep the output linear. The performance is bandwidth limited by the delay of the feedback loop, which must be kept much less than one symbol period. Thus, for a transmission rate of 20 million symbols per second, the symbol period is 50 ns and the allowable delay becomes comparable to that of most microwave systems. Reductions of up

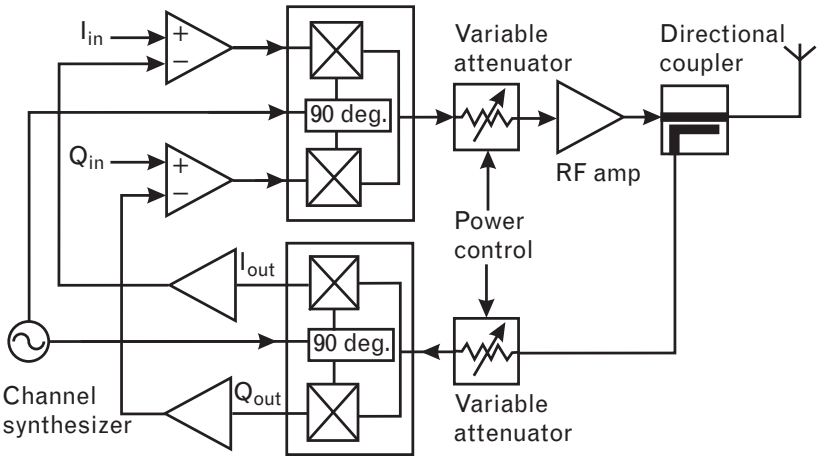


FIGURE 5.76
A Cartesian loop transmitter for linearization of the output signal. (From: [19]. © 2001 Artech House, Inc. Reprinted with permission)

to 35 dB in ACPR have been reported for $\pi/4$-QPSK systems, much more than achievable with other techniques.

Envelope restoration techniques, such as the LINC amplifier [18] enable the use of highly efficient class-D and class-E switching amplifiers to achieve linear amplification and high efficiencies over a broad dynamic range. Because such amplifiers cannot track amplitude variation, DSPs are used to split the input signal $s(t)$ into two components $S1(t)$ and $S2(t)$, as shown in Figure 5.77. As long as the original signal $s(t)$ is bounded in amplitude, as its phase rotates the two components will always lie on the unit circle. Only the phase between them alters in such a way that their sum reconstitutes the original signal, in both amplitude and phase. Because the components are constant in amplitude, the output insensitivity of the switch-mode amplifiers to input signal amplitude is irrelevant as the original signal envelope can be recreated by summing the two output signals. This assumes, however, that the two amplifiers are accurately phase- and gain-matched to each other. Typically, 0.5 dB in gain matching and less than 2° in phase matching are required. Furthermore, the difference signal used to create $S1(t)$ and $S2(t)$ is not narrowband and its spectrum extends far into adjacent channels. Because wideband matching is difficult to achieve with amplifiers in compression, a DSP is used in a closed-loop

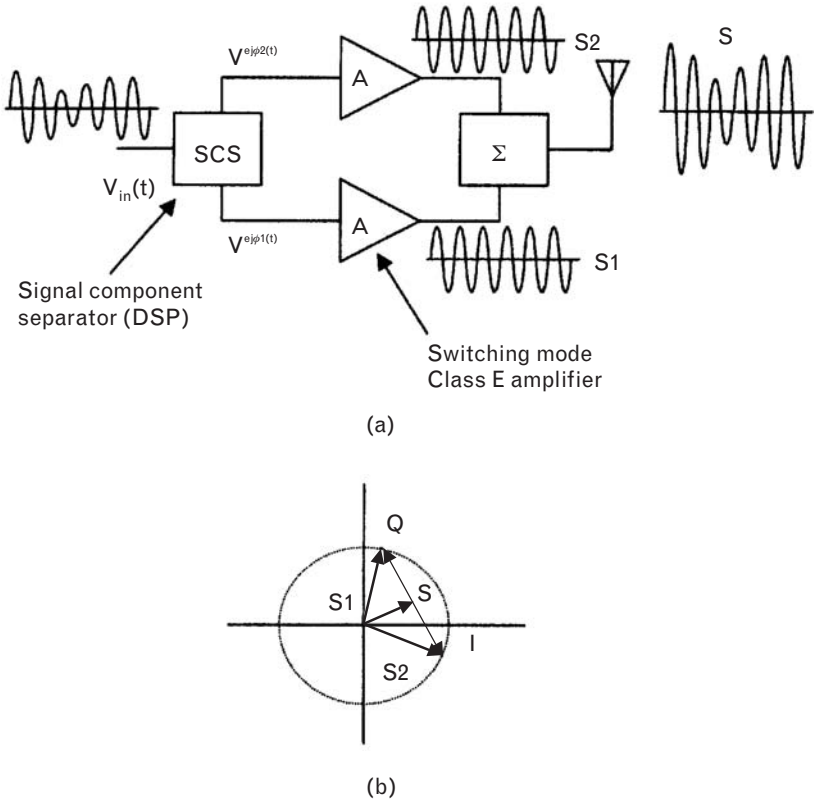


FIGURE 5.77
(a) Schematic architecture of the LINC amplifier. (b) Decomposition of the signal with time-varying envelope into two components with constant envelope. (From: [18]. © 2001 IEEE. Used with permission.)

system to correct the phase of the input signals to compensate for any errors [41]. The use of DSPs in linearizers is increasing, as they provide accurate and even adaptive correction over a wide dynamic range and do not require monotonic distortion behavior [42].

5.7 Problems

1. Redraw the I-V curves for the ideal device of Figure 5.8, and assume I_{DSS} = 500 mA. Show the optimum load line when the device is biased at 10-V drain voltage. What is the optimum load resistance, neglecting V_{SAT}? What is the theoretical saturated output power? Assume that a matching network can be designed to transform the 50-Ω load into this resistance at the intrinsic device terminals. Suppose that after the design is built, we find the saturated output power is not as high as predicted.
   (a) We find that when we increase the supply voltage from 10V to 12V, we can increase the input power further and obtain a higher saturated output power. Is the device voltage-limited or current-limited? Should we increase or decrease the effective resistance at the device terminals? Draw the actual load line to illustrate what is happening.
   (b) Now assume that the device breakdown voltage is 20V. This time, when we increase the supply voltage, the output power decreases. Draw the actual load line to illustrate the most likely scenario.
   (c) We find that when we increase the supply voltage from 10V to 12V, we can increase the input power further but nothing happens to the output power, just as with the previous bias condition. Is the device voltage-limited or current-limited? Should we increase or decrease the effective resistance at the device terminals? Draw the actual load line to illustrate what is happening.
2. Figure 5.78 shows a simple FET amplifier in a 50-Ω measurement system. The bias is brought in through RF chokes, small parasitic resistances are shown at the gate and drain, and blocking capacitors protect the input and output. Load this circuit into a CAD program and observe the effect of changing the gate bias, drain bias, load resistance, and input power on the amplifier load line and the output power, distortion, and efficiency. As a reference point, set the gate bias to yield a quiescent drain current of $I_{DSS}/2$, a drain bias well into the flat portion of the I-V curves, a load resistance approximately equal to R_{LOPT}, and input power low enough to avoid both cutoff and turn-on. The load line should look similar to that in Figure 5.78.

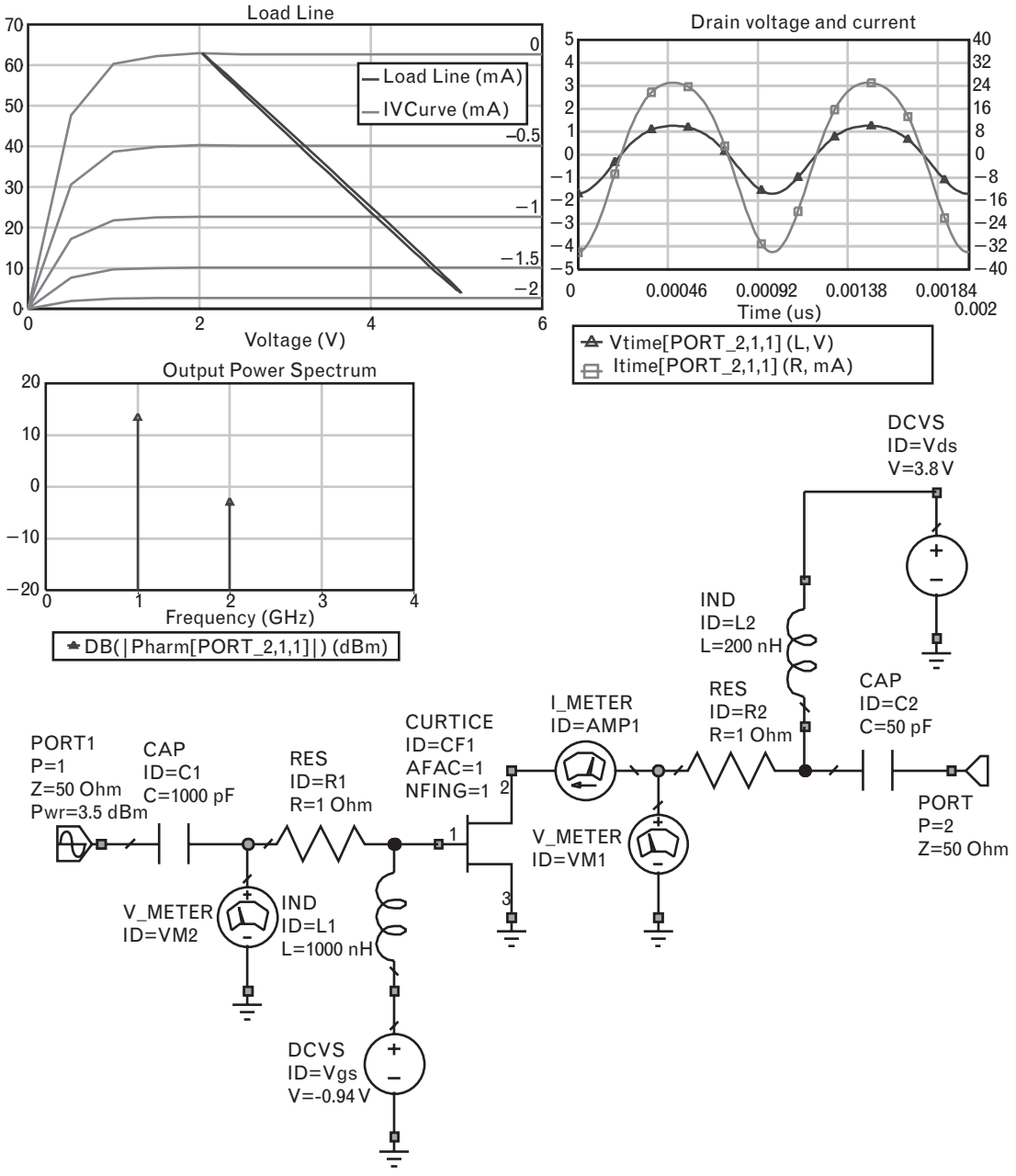


FIGURE 5.78 *Simple FET amplifier and its load line.*

(a) Examine the output spectrum. From the I-V curve, what is the principal cause of the harmonics?
(b) Increase the input power into saturation. What happens to the relative spectral content of each harmonic? From the I-V curve,

what additional effects are causing the harmonics to change? Tune the gate voltage so the second-harmonic is reduced and the third-harmonic is stronger in power. Examine the drain voltage and drain current. What do you observe? What causes the waveforms to take this shape? What else can you tune to increase this effect and create a limiting amplifier with a quick transition from linear to limiting operation?

(c) Return to the reference point. Increase the load resistor, and if necessary the input power so the device voltage is now driven into the knee of the curve. What happens to the distortion (examine the spectral harmonic content)? What mechanism is causing this? What happens to the peaks of the drain voltage and current waveforms? What happens if the drain bias is increased? Why?

(d) Return to the reference point. Decrease the load resistor, and if necessary adjust the input power so the device is just driven between zero and I_{DSS}. What is the effect of increasing the bias voltage? Why?

3. Consider the locus for maximum power from 900 to 1,900 MHz, as shown on the Smith chart in Figure 5.16(b). Try to construct a matching network so that a 50-Ω load follows the locus. Draw the loci for the following, which might at first glance appear to track such an optimum locus with frequency: (a) 50Ω followed by a step-down transformer and a shunt capacitor; (b) 50Ω followed by a step-down transformer and a series capacitor; (c) 50Ω followed by a series capacitor and shunt capacitor; and (d) 50Ω followed by a shunt capacitor and series capacitor.

4. Construct the –2-dB load pull power contour of a device that is biased at 3V and has a maximum current of 60 mA. First neglect the parasitics, then calculate the effect of a 4-pF shunt capacitance and 3-nH series inductor at 1 GHz to transform the optimum load to the external reference planes of the device. Draw the same load pull contour at 2 GHz. What is the direction of the optimum load impedance with frequency? What output matching network can you use to match a 50-Ω load to achieve 2-dB degraded power across 1 to 2 GHz?

5. In Figure 5.23, derive an expression showing why the drain voltage must be backed off by a voltage $V_p/2$ compared with class-A operation, when the signal swing approaches the breakdown voltage. Is this problem a concern with bipolar transistors as well?

6. A single-ended to differential transformer (balun) for a push-pull amplifier can be made with a quarter-wave coaxial line suspended above a ground plane. The shield is grounded at the single-ended input; the output signals are taken from the center conductor and

the shield at the remote end [see Figure 5.26(a)]. Use a circuit simulator to derive the bandwidth, phase, and coupling: (a) for an electrical model at 1 GHz; and (b) for a physical model at 1 GHz on a substrate material of dielectric constant 10. Use the parameters for miniature copper coaxial cable; and (c) as for (b), but wrapping the coaxial cable around a ferrite core.

7. Derive the differences in gain, output power, and efficiency between the class-A amplifier and the harmonic control amplifier of Figure 5.34 when the input voltage is (a) half-sinusoid and (b) rectangular. In all cases assume the gate swings between zero volts and pinch-off.

8. Derive an expression for the error introduced in a third-order intermodulation distortion measurement, when the measured output IMD is *MIMD* (with respect to the carrier) and the source IMD is *SIMD* (with respect to the carrier). Derive the two extreme cases, one where the voltages add in phase, the other where they add out of phase. Does it make any difference if *SIMD* is measured at the input or output (assuming linear amplification)? If the source IMD is 10 dB below the measured IMD from the device under test, what are the bounds on the measurement error?

9. When negative feedback is applied around a device, the gain drops and the output distortion is reduced by the amount of the gain reduction when the fundamental output power is readjusted to the same level as before (by increasing the input power).
   (a) Prove that if an amplifier has an open-loop gain A and a voltage V_0 is fed back and added to the input, the closed-loop gain will be the open-loop gain reduced by $(1 + A)$.
   (b) Differentiate the expression for closed-loop gain and show that the fractional change in closed-loop gain to changes in the open-loop gain is reduced by the same factor. Explain why the distortion is reduced by this factor.
   (c) Derive an expression for $(1 + A)$ in terms of an emitter degeneration resistor inserted in the emitter leg of a transistor.
   (d) Prove that the output third-order intercept point increases by 4/3 times the gain reduction (in decibels), assuming a third-order device nonlinearity.

10. Model the simple diode predistorter of Figure 5.66 using a harmonic balance simulator. Assume a diode model with $I_s = 10^{-13}$ A and $R_b = 200\Omega$, $V_{\alpha} = 0.7$V, and a range of input powers from 0 to 25 dBm. Draw the I-V curve and the associated dc load line of the diode.
    (a) What is the effect of increasing the bias voltage to 0.8V?
    (b) What happens if the bias resistor is decreased to 50Ω? What

changes to the input match must be made?
(c) What happens in both cases as the diode is overdriven and starts to rectify the input signal? Plot the current and voltage waveforms in the diode for small-signal and large-signal input power levels.

11. Consider the small-signal model of a dual-gate FET as a cascode connection of two FETs, and its use as a predistorter to linearize a signal.
(a) Using the small-signal model of a common source FET shown in Figure 5.79, show that if the output conductance G_{d1} decreases with increasing input power, the phase deviation of the output voltage will be positive with increasing input power, assuming the phase of input voltage v_1 remains constant. Assume for this part of the problem that the loading of the second FET is a constant load impedance Z_L.
(b) Now show that for increasing input voltage v_3 on the common-gate FET, a decrease in G_{d2} causes a negative phase deviation of the output voltage v_2.
(c) Derive the condition for which the phase deviations can be made equal and opposite.

12. In this problem, set up a bipolar transistor amplifier for class-AB operation in a 50-Ω system in a nonlinear simulator. Set the small-signal quiescent current to 10% of I_{MAX}. For each of the cases below, increase the input power so that the steady-state dc collector current increases to 50% of I_{MAX}, and record the two-tone fundamental and third-order distortion output power: (a) with a dc voltage source directly connected to the base; (b) with a dc current source directly connected to the base; and (c) with a resistive divider network at the base with Thevenin equivalent bias resistor equal to 100Ω. How do the results compare with those of Figure 5.74(b)? What difficulties did you encounter in simulating the problem? Were there any unexpected side effects of modifying the bias conditions?

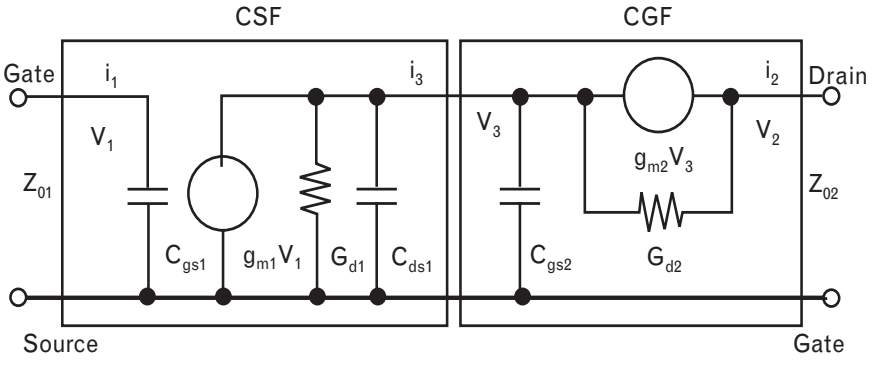


Figure 5.79
A simplified small-signal model for the common-source FET followed by the common-gate FET, used for self-phase distortion compensation. (From: [27]. © 1995 IEEE. Used with permission.)

13. Consider the amplifier design as built in Section 5.4.3 in Figure 5.48. Load the circuit into the simulator and recreate the conditions discussed.
(a) Remove the RF input power, and adjust the gate bias resistor to 100Ω. What is the difference in gate bias voltage as the drive increases? Is the linearity (measured by a two-tone test) any different?
(b) Perform a power sweep at the gate, increasing the input power from 10 to 30 dBm in 1-dB steps. What happens as the number of harmonics is changed from 4 to 6 to 8 to 10? What is the minimum number of harmonics required for accurate simulation?
(c) Reduce the quiescent drain current to 100 mA with no drive (class-AB). What happens when nominal input power is applied? Why? Plot the variation of dc drain current when the drive is increased to 30 dBm.
(d) Increase the drain bias voltage to 6V and maintain 800-mA quiescent current. Calculate the expected output power and optimum load resistance. Reoptimize for maximum output power, and compare with your calculations. Investigate and explain any differences.
(e) Examine the second and third-harmonic impedances of the actual load circuit. How might you improve the output match to get more power and better efficiency?

14. With a nonlinear simulator, recreate the circuit of Figure 5.48 but replace the input and output matching networks with load pull tuners. Set the input tuner to a reflection coefficient of 0.95, 192°, and the output tuner to 0.88, 192°. Vary the gate bias at drain voltages of 3V and 6V, and examine the gain from input power levels of 15 to 30 dBm.
(a) What is the gate bias that makes the amplifier closest to an ideal limiter? (*Note:* An ideal limiter has $P_{1dB} \approx P_{SAT}$ and constant P_{SAT} with input power). What makes this amplifier closest to an ideal limiter at this bias level? (*Hint:* Examine the fundamental components of current and voltage, and their waveforms, and look for the best square wave behavior in saturation.)
(b) What is the AM/PM conversion observed, in degrees per decibel? Can you think of any way to linearize this characteristic?
(c) From settings for the ideal limiter, shift the bias point into class-B by lowering the gate bias voltage, and repeat the process to illustrate the shortcomings of a soft limiter.
(d) With the original circuit, increase the load resistance at the intrinsic drain to 8Ω (e.g., by terminating the output in 100Ω instead of 50Ω). What is the impact on the waveforms and limiting operation? Is the device now voltage-limited, as expected?

References

[1] Cripps, S. C., "A Theory for the Prediction of GaAs FET Load-Pull Power Contours," *IEEE Intl Microwave Symposium Digest*, Boston, MA, 1983, pp. 221–223.

[2] Clarke, K., and D. Hess, *Communication Circuits Analysis and Design*, Reading, MA: Addison-Wesley, 1971.

[3] Vendelin, G. D., A. M. Pavio, and U. L. Rohde, *Microwave Circuit Design Using Linear and Nonlinear Techniques*, New York: John Wiley & Sons, 1990.

[4] Sevick, J., *Transmission Line Transformers*, Atlanta, GA: Noble Publishing, 1996.

[5] Duvanaud, C., et al., "High-Efficient Class F GaAs FET Amplifiers Operating with Very Low Bias Voltages for Use in Mobile Telephones at 1.75 GHz," *IEEE Microwave and Guided Wave Letters*, Vol. 3, No. 8, August 1993, pp. 268–270.

[6] Cripps, S., *RF Power Amplifiers for Wireless Communications*, Norwood, MA: Artech House, 1999.

[7] Ingruber, B., "Reliability and Efficiency Aspects of Harmonic-Control Amplifiers," *IEEE Microwave and Guided Wave Letters*, Vol. 9, No. 11, November 1999.

[8] Ingruber, B., et al., "Rectangularly Driven Class-A Harmonic-Control Amplifier," *IEEE Trans. on Microwave Theory and Techniques*, Vol. 46, No. 11, November 1998, pp. 1667–1672.

[9] Raab, F., "Class-E, Class-C, and Class-F Power Amplifiers Based Upon a Finite Number of Harmonics," *IEEE Trans. on Microwave Theory and Techniques*, Vol. 49, No. 8, August 2001, pp. 1462–1468.

[10] Kobayashi, H., J. Hinrichs, and P. Asbeck, "Current-Mode Class-D Power Amplifiers for High-Efficiency RF Applications," *IEEE Trans. on Microwave Theory and Techniques*, Vol. 49, No. 12, December 2001, pp. 2480–2483.

[11] Raab, F., "Class-F Power Amplifiers with Maximally Flat Waveforms," *IEEE Trans. on Microwave Theory and Techniques*, Vol. 45, No. 11, November 1997, pp. 2007–2012.

[12] Ortega-Gonzalez, F., et al., "High-Efficiency Load-Pull Harmonic Controlled Class-E Power Amplifier," *IEEE Microwave and Guided Wave Letters*, Vol. 8, No. 10, October 1998.

[13] Mader, T., and Z. Popovic, "The Transmission-Line High-Efficiency Class-E Amplifier," *IEEE Microwave and Guided Wave Letters*, Vol. 5, No. 9, September 1995.

[14] Martin, A. L., and A. Mortazawi, "A Class-E Power Amplifier Based on an Extended Resonance Technique," *IEEE Trans. on Microwave Theory and Techniques*, Vol. 48, No. 1, January 2000, pp. 95–96.

[15] Wilkinson, A., and J. Everard, "Transmission-Line Load-Network Topology for Class-E Power Amplifiers," *IEEE Trans. on Microwave Theory and Techniques*, Vol. 49, No. 6, June 2001, pp. 1202–1209.

[16] Piper, I., et al., "Balanced LNA Suits Cellular Base Station," *Microwaves and RF,* April 2002.

[17] De Carvalho, N. B., and J. C. Pedro, "A Comprehensive Explanation of Distortion Sideband Asymmetries," *IEEE Trans. on Microwave Theory and Techniques*, Vol. 50, No. 9, September 2002, pp. 2090–2101.

[18] Asbeck, P. M., L. E. Larson, and I. G. Galton, "Synergistic Design of DSP and Power Amplifiers for Wireless Communications," *IEEE Trans. on Microwave Theory and Techniques*, Vol. 49, No. 11, November 2001, pp. 2163–2169.

[19] Kennington, P., *High-Linearity RF Amplifier Design*, Norwood, MA: Artech House, 2001.

[20] De Carvalho, N. B., and J. C. Pedro, "Compact Formulas to Relate ACPR and NPR to Two-Tone IMR and IP3," *Microwave Journal*, December 1999, pp. 70–84.

[21] Falconer, D., et al., "Frequency-Domain Equalization for Single-Carrier Broadband Wireless Systems," *IEEE Communications Magazine*, April 2002, pp. 58–66.

[22] Struhsaker, P., and K. Griffin, "Analysis of PHY Waveform Peak to Mean Ratio and Impact on RF Amplification," IEEE 802.16a cont. IEEE 802.16.3c-01/46, March 6, 2001.

[23] Leenaerts, D., J. van der Tang, and C. Vaucher, *Circuit Design for RF Transceivers*, Boston, MA: Kluwer Academic Publishers, 2001.

[24] Yamauchi, K., et al., "A Microwave Miniaturized Linearizer Using a Parallel Diode with a Bias Feed Resistance," *IEEE Trans. on Microwave Theory and Techniques*, Vol. 45, No. 12, December 1997, pp. 2431–2433.

[25] Kim, M., C. Kim, and H. Yu, "An FET-Level Linearization Method Using a Predistortion Branch FET," *IEEE Microwave and Guided Wave Letters*, Vol. 9, No. 6, June 1999, pp 233–235.

[26] Haskins, C., T. Winslow, and S. Raman "FET Diode Linearizer Optimization for Amplifier Predistortion in Digital Radios," *IEEE Microwave and Guided Wave Letters*, Vol. 10, No. 1, January 2000, pp. 21–23.

[27] Hayashi, H., M. Nakatsugawa, and M. Muraguchi, "Quasi-Linear Amplification Using Self Phase Distortion Compensation Technique," *IEEE Trans. on Microwave Theory and Techniques*, Vol. 43, No. 11, November 1995, pp. 2557–2564.

[28] Kim, J., et al., "A New 'Active' Predistorter with High Gain and Programmable Gain and Phase Characteristics Using Cascode-FET Structures," *IEEE Trans. on Microwave Theory and Techniques*, Vol. 50, No. 11, November 2002, pp. 2459–2466.

[29] Jing, D., et al., "New Linearization Method Using Interstage Second Harmonic Enhancement," *IEEE Microwave and Guided Wave Letters*, Vol. 8, November 1998, pp. 402–404.

[30] Samelis, A., and D. Pavlidis, "Mechanisms Determining Third Order Intermodulation Distortion in AlGaAS/GaAs Heterojunction Bipolar Transistors," *IEEE Trans. on Microwave Theory and Techniques*, Vol. 40, No. 12, December 1992, pp. 2374–2380.

[31] Maas, S. A., B. L. Nelson, and D. L. Tait, "Intermodulation in HBTs," *IEEE Trans. on Microwave Theory and Techniques*, Vol. 40, No. 3, March 1992, pp. 442–448.

[32] Kim, W., et al., "Analysis of Nonlinear Behavior of Power HBTs," *IEEE Trans. on Microwave Theory and Techniques*, Vol. 50, No. 7, July 2002, pp. 1714–1721.

[33] Bailey, M. J., "Intermodulation Distortion in Pseudomorphic HEMTs and an Extension of the Classical Theory," *IEEE Trans. on Microwave Theory and Techniques*, Vol. 48, No. 1, January 2000, pp. 104–110.

[34] Abrie, P. L., *Design of RF Microwave Amplifiers and Oscillators*, Norwood, MA: Artech House, 2000.

[35] Teeter, D. A., J. R. East, and G. I. Haddad, "Use of Self-Bias to Improve Power Saturation and Intermodulation Distortion in CW Class-B HBT Operation," *IEEE Microwave and Guided Wave Letters*, Vol. 2, No. 5, May 1992, pp. 174–176.

[36] Yang, Y., and B. Kim, "A New Linear Amplifier Using Low-Frequency Second-Order Intermodulation Component Feedforwarding," *IEEE Microwave and Guided Wave Letters*, Vol. 9, No. 10, October 1999, pp. 419–421.

[37] Gilmore, R. J., R. Kiehne, and F. J. Rosenbaum, "Circuit Design to Reduce Third-Order Intermodulation Distortion in Microwave MESFET Amplifiers," *IEEE 1985 International Microwave Symposium Digest*, June 1985.

[38] Iwamoto, M., et al., "An Extended Doherty Amplifier with High-Efficiency Over a Wide Power Range," *IEEE Trans. on Microwave Theory and Techniques*, Vol. 49, No. 12, December 2001, pp. 2472–2475.

[39] Kennington, P., "Linearized Transmitters: An Enabling Technology for Software Defined Radio," *IEEE Communications Magazine*, February 2002, pp. 156–162.

[40] Kennington, P., "Methods Linearize RF Transmitters and Power Amps," *Microwaves and RF*, December 1998.

[41] Zhang, X., et al., "Gain/Phase Imbalance-Minimization Techniques for LINC Transmitters," *IEEE Trans. on Microwave Theory and Techniques*, Vol. 49, No. 12, December 2001, pp. 2507–2510.

[42] Katz, A., "Linearization: Reducing Distortion in Power Amplifiers," *IEEE Microwave Magazine*, December 2001, pp. 37–49.

CHAPTER 6

Oscillators

Oscillators have often been regarded as a black art within an industry that has long held that reputation as a whole. That perception need no longer be the case, because modern CAD tools now not only enable an oscillator to be analyzed in its nonlinear entirety, but even noise to be modeled as part of the design process. In this chapter, we will explore the principles of oscillator design, introducing a variety of techniques and applying these to several examples.

Oscillators are typically characterized as either L-C or R-C oscillators. We will work exclusively with L-C, or resonant, oscillators. L-C oscillators use a resonant circuit modeled by an inductor and capacitor to set the oscillation frequency. Depending on the frequency range, the resonant circuit is realized using a crystal (up to 500 MHz), dielectric transmission lines (500 MHz to 5 GHz), or dielectric resonators (2 to 40 GHz). The other major category of oscillator, the R-C type, is more commonly found in integrated circuits. Lacking inductors, R-C oscillators have much lower Q than L-C oscillators, and instead rely upon the charging and discharging of a capacitor to reach a threshold voltage that causes switching from one mode to another (as in the relaxation oscillator), or rely upon the propagation time delay and inversion through several devices to achieve a delayed output that can be fed back to the input (as in the ring oscillator). Although R-C oscillators are noisier than L-C oscillators, they can be tuned over much larger bandwidths (up to a decade), simply because the charging resistance can be implemented using an active device whose impedance can then be varied over a large range.

Ideally, an oscillator will generate an output current of the form

$$i(t) = A\cos(\omega_0 t) = A\cos(2\pi f_0 t)$$

This is a pure sinusoid, represented by a single phasor of frequency f_0 in the frequency domain. In practice, both A and f_0 will fluctuate about their average values. The first fluctuation is amplitude noise and is generally lower in power than the second fluctuation, known as phase noise. Achieving the desired levels of A and f_0, minimizing the sources of phase noise, and tuning the frequency f_0 are the key oscillator design criteria that we will consider in this chapter.

6.1 Principles of oscillator design

We begin our study of oscillators with the classical control theory approach. Although less commonly used in practical circuit design, it is a useful starting point, because this theory enables us to design an oscillator as a two-port component, in which the input-output response can be determined, and the conditions for oscillation exactly derived. It turns out to be very powerful for synthesis of oscillators because cause and effect can be examined over a broad frequency range. However, we quickly move on to consider oscillators as components with a single port (i.e., the output port), and we will review the large body of theory associated with one-port oscillator design. The one-port approach is a more common approach for oscillator design at RF frequencies, so later in the chapter we will illustrate its power with several examples.

6.1.1 Two-port oscillator design approach

A two-port device has the decided advantage of having both an input and an output port. Although it may seem strange that such an approach can be used when an oscillator has only a single output port, the oscillator can, in fact, be modeled as an open-loop system in which one port has disappeared because the output has been fed back to the input. As a result of closing the loop, both the output port and input port are subsumed into the oscillator circuit and disappear. We must be careful to distinguish the meaning of the term "output" from the context—sometimes it refers to the *output* of the open-loop system, or sometimes to the port at which the oscillator load is connected and from which we take *output* power. We will see that the two need not necessarily be the same.

6.1.1.1 Closed-loop system analysis of an oscillator

Figure 6.1 shows an example of the closed-loop system we shall consider. The amplifier of gain A represents the linear gain of the transistor, and the gain block $L(s)$ its limiting characteristics as the device begins to be driven with strong enough signal levels to compress.[1] The feedback component is a linear filter of some sort represented by its transfer characteristic $H(s)$. In this system, the input voltage $V_I(s)$ is either thermal noise or a step response generated when bias is applied to the device. This voltage will be removed in steady state, but it is required to simulate startup of the oscillator. Control theory states that this network must have a pair of complex conjugate poles in the right-half complex plane to commence sinusoidal oscillation that increases with an exponentially growing envelope. For steady-state

1. The s is the Laplace transform variable that is used in systems analysis where $s = j\omega$.

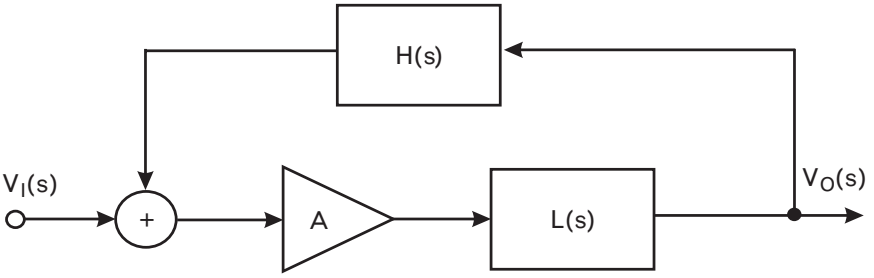


FIGURE 6.1
Closed-loop system for modeling the feedback of an oscillator.

oscillation, the action of the limiter block must be to shift those poles towards the imaginary axis, so that at steady state, a constant amplitude waveform results at the frequency given by the location of the poles on the imaginary axis. In theory, this requires knowledge of the root locus plot[2] of the system and its behavior as the gain of the system $AL(s)$ changes, but in circuit design practice at least, this is rarely necessary.

Figure 6.2 shows an example of an oscillator in which the feedback loop of Figure 6.1 is clearly identifiable. One of the difficulties of the two-port approach is that it is sometimes very difficult at RF frequencies to identify the feedback path—in fact, for some oscillators, the feedback path is through the device itself, so cannot be explicitly represented in a circuit diagram. However, for the Hartley oscillator shown in Figure 6.2, the feedback path $H(s)$ is evident through the autotransformer. The device itself provides both the small-signal gain A and the limiting function $L(s)$. There is no explicit input port nor applied input voltage because the circuit

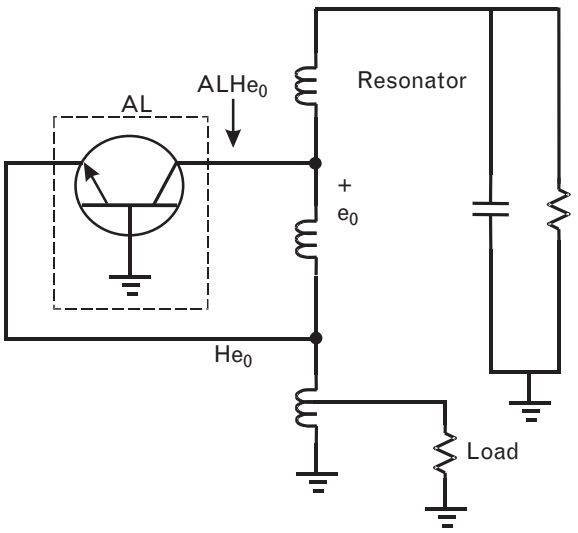


FIGURE 6.2
A Hartley oscillator, showing the feedback path between the input (emitter) and output (collector).

2. For those that have forgotten, the root locus is a plot of the poles of $1 - AL(s)H(s)$ in the $s = \sigma + j\omega$ plane as the gain $AL(s)$ of the system (usually a function of g_m) increases. The poles represent values of s for which the closed-loop gain (6.2) becomes infinite.

is closed loop. However, the emitter could be considered as an "input" for purposes of analysis, and the device noise voltage modeled at the emitter could be considered additive to the feedback signal and provide the initial excitation of the oscillator.

It is clear from Figure 6.1 that the output voltage is given by

$$V_O(s) = AL(s)\big(V_I(s) + H(s)V_O(s)\big) \tag{6.1}$$

so that solving for the ratio of output to input we obtain

$$\frac{V_O(s)}{V_I(s)} = \frac{AL(s)}{1 - AL(s)H(s)} \tag{6.2}$$

The expression in the numerator is just the *forward* gain of the system, and the expression $AL(s)H(s)$ is the total *loop* gain of the oscillator. The poles of the system are given by those values of the complex frequency s for which

$$1 - AL(s)H(s) = 0 \tag{6.3}$$

We shall examine the conditions for startup of oscillation in more detail in a later section, but for steady-state oscillation, the roots of the equation must lie on the imaginary axis at a frequency $s = j\omega$ for which

$$\begin{aligned} \Re e\big(AL(j\omega_0)H(j\omega_0)\big) &= 1 \\ \Im m\big(AL(j\omega_0)H(j\omega_0)\big) &= 0 \end{aligned} \tag{6.4}$$

Equations (6.4) provide a *necessary* (but not sufficient) condition for oscillation, and they are known as the Barkhausen criterion [1]. These two equations can be solved for the signal level that results in the limiter assuming the value necessary to meet (6.4) at the frequency ω_0. These occur when the total loop gain is equal to one, and the total phase shift around the loop equals zero, or a multiple of 2π radians.

It is customary to assume that for startup of oscillations the loop gain must be greater than one, so that as the transconductance g_m of the transistor starts to reduce as the signal levels grow, and the limiter starts to limit, the Barkhausen criterion is satisfied at steady state (i.e., loop gain is exactly equal to one) at a reasonable amplitude of oscillation. For many oscillator circuits (those with a pair of complex conjugate poles and a zero at the origin), it can be shown [2] that this corresponds to $g_m > G_L$ at small signals, where G_L is the total load conductance seen at the current source intrinsic to the device output. Equation (6.4) can therefore generally be met at

steady state by designing the system to ensure loop gain greater than one at small-signal levels with net loop phase shift equal to zero. This is usually, but unfortunately not always, the case. In a later section, we will show instances of circuits that satisfy (6.4) but are unable to start oscillating at small-signal levels because the poles of the root locus do not lie in the right half plane at small-signal levels.

6.1.1.2 Examples of open-loop oscillator design

The analysis above suggests a relatively straightforward approach to the design of oscillators. In fact, the starting point to design an oscillator is just to design an amplifier. Suppose we want to design a 1,000-MHz oscillator using the approach.

The circuit shown in Figure 6.3 is a straightforward small-signal amplifier, using the BFP640 HBT packaged transistor that we have used in previous examples. Initially, we have modeled the transistor by its small-signal *S*-parameters at 3-V collector voltage and 30-mA dc current. We have added both resistive and emitter feedback in order to provide a good input and output match for the device, and to flatten the gain with frequency. Series inductances are also used at the input and output as well for matching purposes and also as variables to (later) adjust the phase shift around the loop. In fact, the only unusual aspect of the design is that the resistor that ultimately becomes the oscillator load must be incorporated into the open-loop design from the very beginning. If we wish to design this oscillator at 1,000 MHz, then when the loop is closed, (6.4) must be satisfied to ensure steady-state oscillation. However, when the loop is closed the (open-loop) output is connected back to the input, so these ports disappear and an oscillator load must be specified and a possibly new (oscillator) output port designated. Its associated load resistor needs to be included in the open-loop design from the outset.

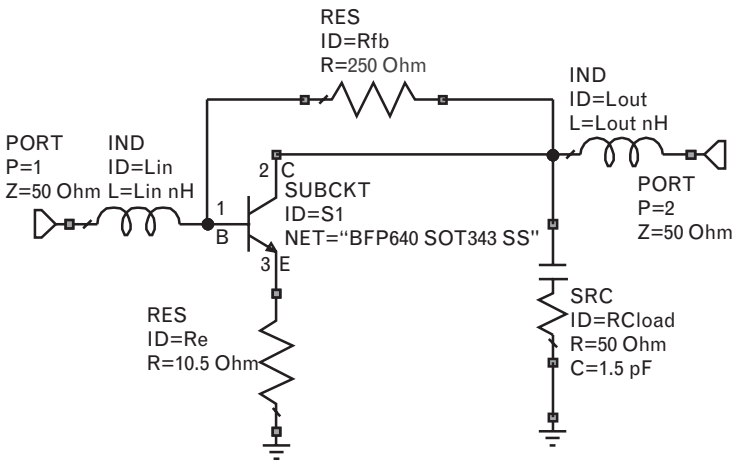


FIGURE 6.3
Small-signal amplifier used as an open-loop system as a starting point for oscillator design.

Even if the oscillator output port remains the same as the open-loop output port, the load impedances from the open-loop measurement ports are effectively replaced by the loading of the amplifier input and output itself under closed-loop conditions. The amplifier design must therefore ensure that the gain still remains unity and the phase around the loop still equals zero at the oscillator signal levels once the loop is closed. One would expect that we can simulate these from the magnitude and phase of the circuit s_{21}. However, because s_{21} is measured in a 50-Ω system, the measurement is only representative of the open-loop gain and phase if the input and output terminating impedances remain 50Ω after the loop is closed. Any reactive loading or mismatch from the input load on the output or the output load on the input after the loop is closed will change both the phase and amplitude of the s_{21} and thus the system gain. Consequently, it is necessary not only to ensure that the open-loop system gain s_{21} meets the Barkhausen conditions (6.4), but also that the input and output match of the small-signal amplifier are excellent so that closing the loop does not change the measurement conditions under which we obtained s_{21}.

The results of the amplifier analysis are shown in Figures 6.4 and 6.5. By tuning the feedback resistors in particular we can obtain a very good match at both input and output and achieve an open-loop gain greater than one, so that in compression the gain will reduce to one and the phase will hopefully remain constant. However, the phase shift in Figure 6.4 is still about 140° rather than zero.

Some optimization of the design is therefore necessary to simultaneously achieve zero phase and positive gain for the open-loop system s_{21} at 1,000 MHz. If we wish the oscillator to be stable at this frequency, the amplifier needs to be cascaded with a resonant circuit that "locks" its phase

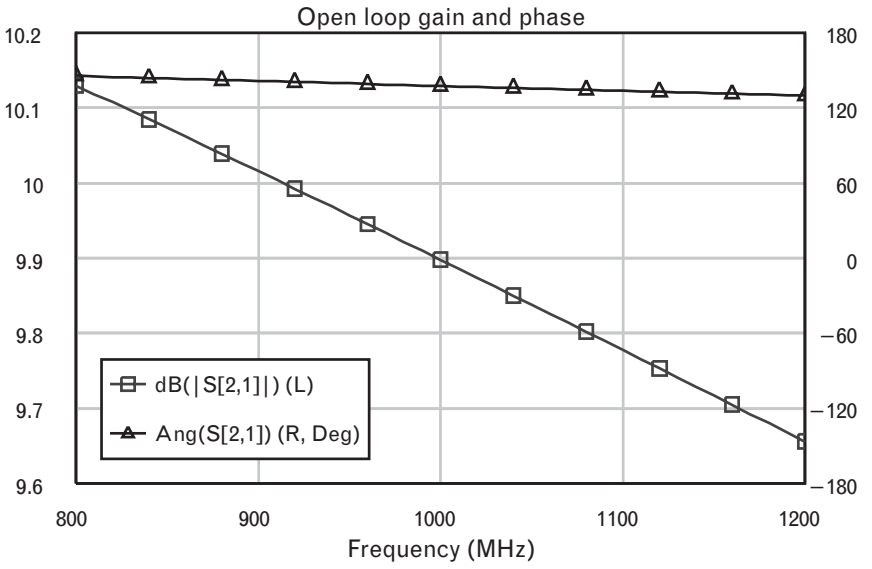


FIGURE 6.4
Open-loop gain and phase of the amplifier of Figure 6.3, taken from its s_{21}.

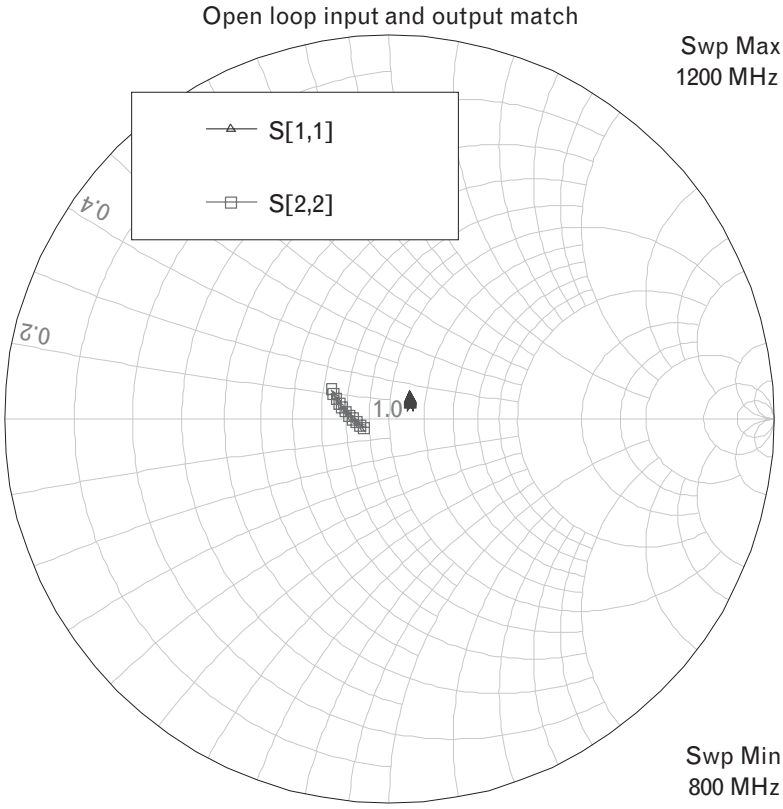


FIGURE 6.5
Open-loop input and output match of the amplifier of Figure 6.3.

to zero tightly around 1,000 MHz. The resulting open-loop amplifier system is shown in Figure 6.6, in which the feedback resistors, series inductors, and resonant circuit are optimized to achieve a good input and output match and to set the total circuit s_{21} equal to +1 at 1,000 MHz.

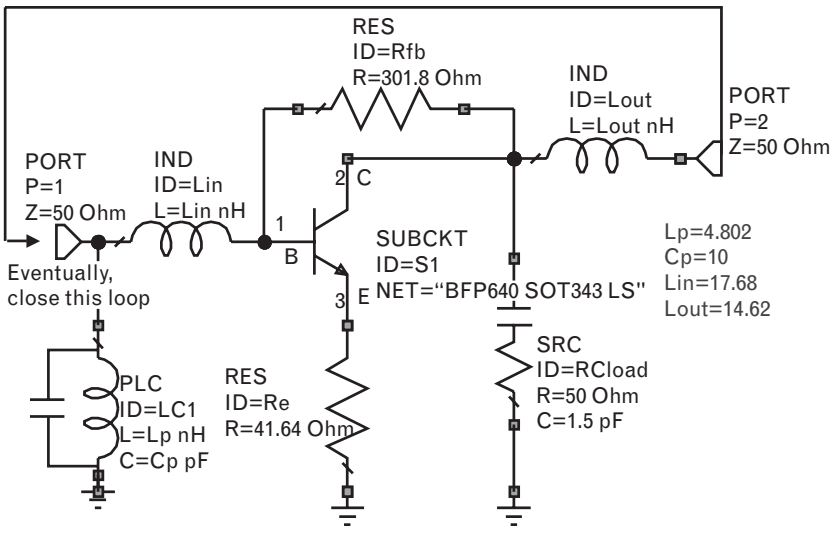


FIGURE 6.6
Open-loop amplifier circuit of Figure 6.3 with an additional resonant section at the input to achieve zero net phase (i.e., positive feedback when the loop is closed as shown).

Unfortunately, this approach is not accurate because we assumed the open-loop system to be unilateral. In fact, we have neglected the effect of the system reverse gain s_{12}, which means that even if the above conditions are met, the closed-loop system will not oscillate at the expected design frequency.

Randall and Hock [3] have derived an expression for the open-loop gain of a system that accounts for the effect of imperfect input and output match, and a nonunilateral amplifier ($s_{12} \neq 0$). Their expression for the open-loop gain is

$$G = \frac{s_{21} - s_{12}}{\left(1 - s_{11}s_{22} + s_{12}s_{21} - 2s_{12}\right)} \tag{6.5}$$

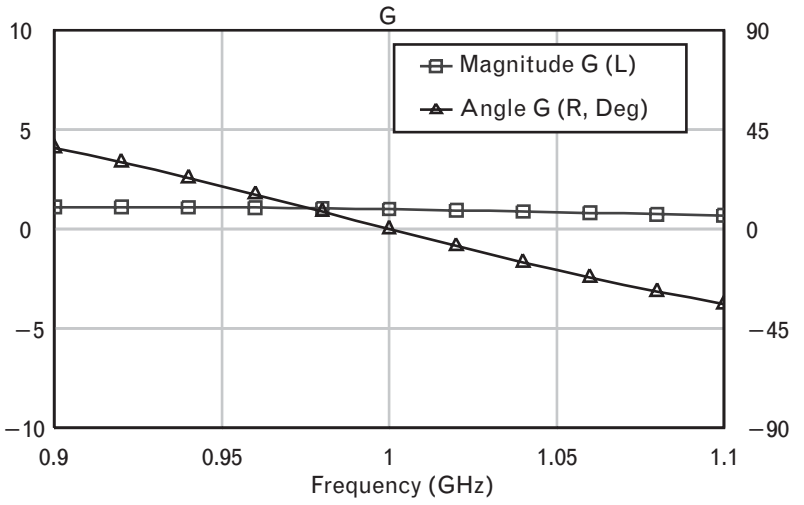
and is derived using an eigenvalue approach of an infinite cascade of such systems whereby the termination impedances and reverse feedback are accounted for. The *S*-parameters are those of the entire open-loop system. Only when the system s_{11} or s_{22} are small, *and* the circuit unilateral ($s_{12} = 0$), does $G = s_{21}$ and will s_{21} alone properly represent the open-loop gain of the system.

Because G accounts for the effects of imperfect termination impedances in the open-loop system, the system needs only to be optimized so that G equals +1 (zero phase) in steady state. This has been done in the circuit of Figure 6.6. In addition, we will now also use large-signal transistor *S*-parameters to calculate the system *S*-parameters in (6.5), to represent the HBT limiting action during steady-state oscillation when the open-loop gain compresses to one.

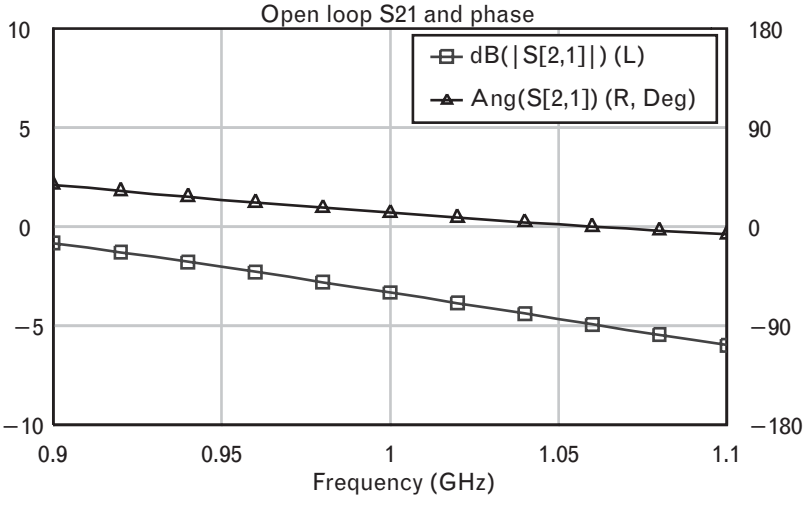
Large-signal transistor *S*-parameters provide a first-order approximation to device behavior over frequency, and they principally model the effect of gain compression as the device is driven large-signal at its input. They can be generated using a nonlinear model for the device, simulating the reflection coefficient and gain when the device is mounted in a 50-Ω system with increasing input power first at the input, and then the output. An alternative, but less accurate approach, is to start with the common-emitter or common-source small-signal *S*-parameters for the device, and to simply multiply the magnitude of the transistor s_{21} by 0.891, 0.794, or 0.708, respectively, to model 1-, 2-, or 3-dB compression. The phase of s_{21} and all the other transistor small-signal *S*-parameters are assumed to remain invariant as the device compresses. This is a rather gross assumption to make, but is still useful to first-order for gaining a good intuitive understanding of circuit operation across frequency.

Here, we take the latter approach, which is simpler. The resulting open-loop responses for the entire circuit are shown in Figures 6.7 and 6.8.

These results are interesting because although the circuit has been optimized so the gain parameter G is exactly +1 at 1,000 MHz, the resulting



(a)



(b)

FIGURE 6.7
(a) Open-loop gain parameter G and (b) the resulting open-loop system s_{21}, after the circuit of Figure 6.6 is optimized for G = +1 (zero phase) at 1,000 MHz.

magnitude of the open-loop system s_{21} is only 0.68 (–3.4 dB) and its phase 12°, rather than 1.0 and 0° as the unwary might expect.[3] The input and output match are also quite poor. However, when the loop is closed as indicated in Figure 6.6, and the oscillator output resistance explicitly "pulled out" to create a new output port for oscillator analysis, we can analyze the circuit of Figure 6.9 at its new output port. The previous 50-Ω resistor is now part of the external measurement system.

3. This has nothing to do with using large-signal *S*-parameters for the transistor to model its gain compression, since *G* and the circuit *S*-parameters for the open-loop system are both calculated using these same large-signal transistor *S*-parameters.

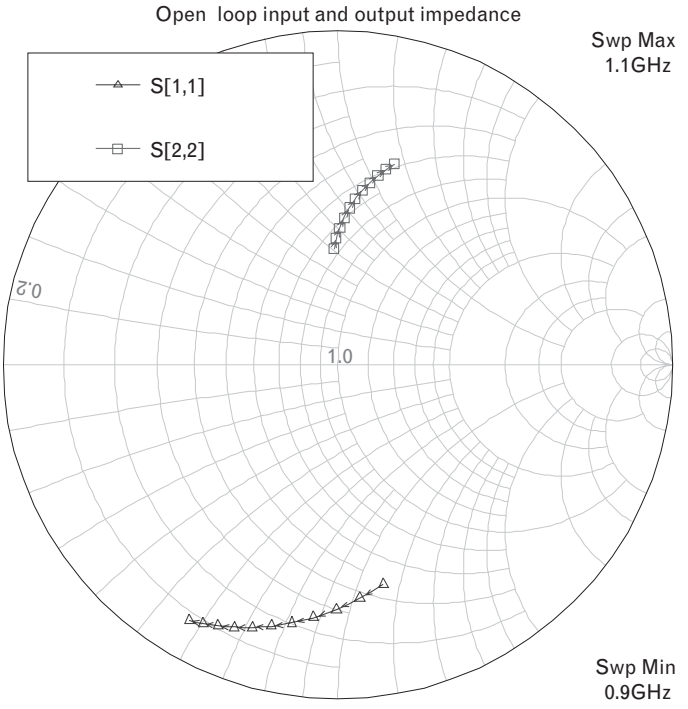


FIGURE 6.8
Resulting input and output match of the circuit of Figure 6.6 after G is optimized to equal +1.

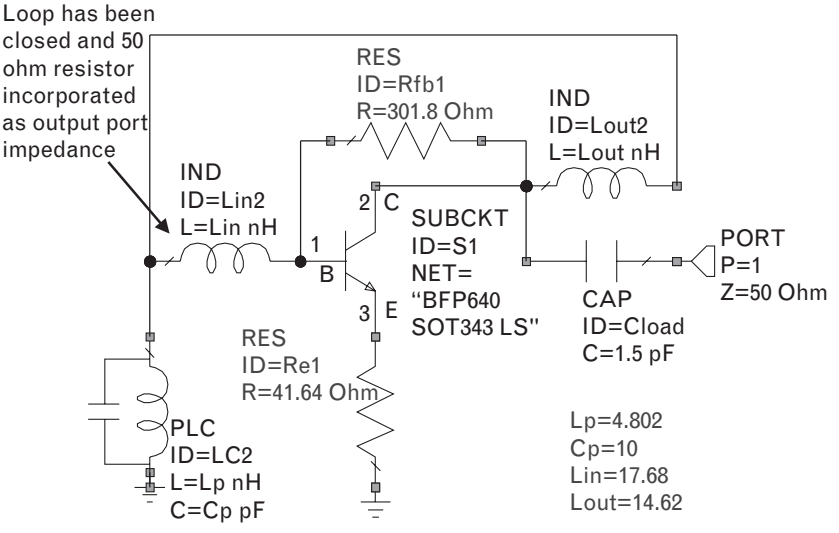


FIGURE 6.9
Oscillator formed by closing the feedback loop in Figure 6.6.

The reflection coefficient as the frequency is swept between 500 and 1,500 MHz is shown in Figure 6.10. For frequencies between about 850 and 1,100 MHz, the reflection coefficient is greater than one and falls outside the "normal" Smith chart.

This is confirmed in Figure 6.11, which shows that the real part of the oscillator impedance is negative between these frequencies. At 1,000 MHz, the input resistance looking into the one-port is exactly –50Ω, equal and

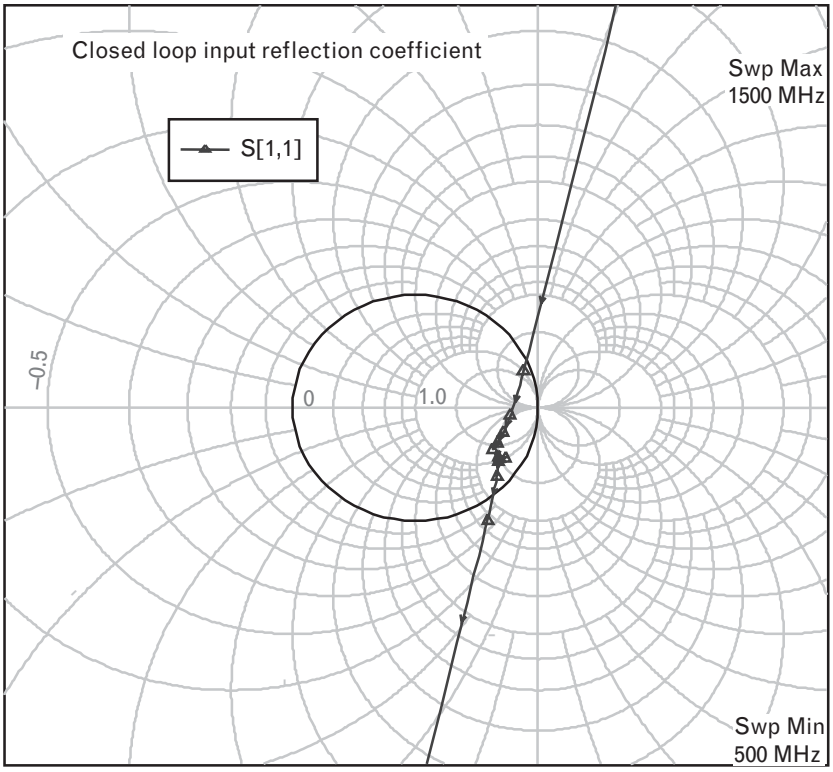


FIGURE 6.10
Reflection coefficient of the oscillator in Figure 6.9, measured at its output port.

opposite to the load resistance placed there. (If we had optimized so that s_{21} = 1 instead of *G*, we would not get –50Ω. Try it!)

Such a negative resistance at the output port of an oscillator is typical. With oscillators, we will see in the next section that a consequence of creating complex poles in the right half-plane is that the real part of the

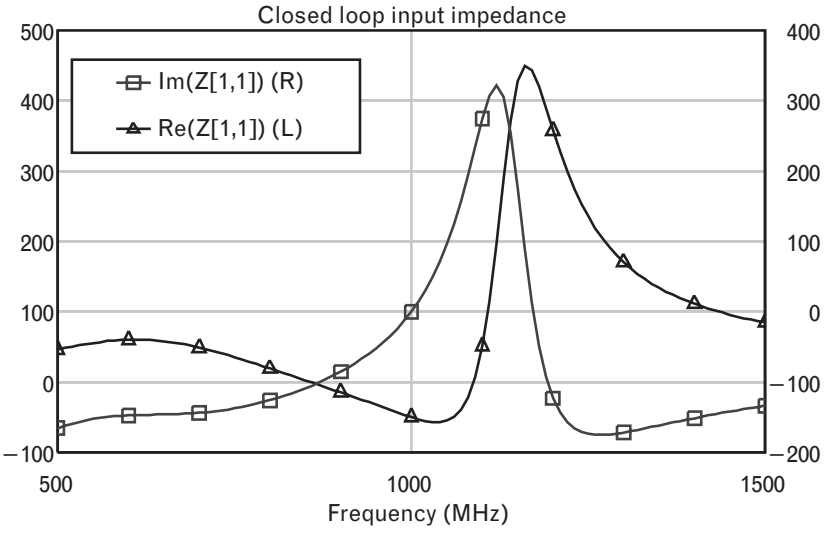


FIGURE 6.11
Real and imaginary parts of the impedance looking into the output port of the oscillator in Figure 6.9.

impedance is negative. This suggests a totally different way of studying oscillators, through one-port analysis, where the design principle is to embed the transistor into a circuit that creates a negative resistance seen looking into the output of the device.

The following sections of this chapter will explore some other concerns in oscillator design, and in particular how to more tightly lock the oscillation frequency or reduce the phase noise by increasing the oscillator Q factor. We did not account for these considerations in the design of our open-loop system, but these should be revisited in this design if they are of concern. For example, the strong resonance in the impedance characteristic at 1,170 MHz is not at all set by the resonant frequency of the added shunt L-C circuit but by the phase of the device itself. This resonance could be shifted to other frequencies by tuning an added series transmission line at the collector, to offset the phase of the device s_{21}. In this example, the shunt L-C resonant circuit was, in fact, tuned by the CAD optimizer to resonate around 450 MHz in order to achieve $G = +1$ at the desired oscillation frequency, a fact reflected in the sharp change in G around 725 MHz. Ideally, we would set the frequency of the resonator close to 1,000 MHz, and optimize other circuit variables instead in order to achieve $G = +1$.

A second example of open-loop design is given in [4]. An FET amplifier was constructed as shown in Figure 6.12.

This amplifier was constructed to have nominal open-loop gain of 7 dB, and was designed with a coupler and delay line combination that feeds about one-quarter of the amplifier output power back to the input. The length of the delay line was chosen for zero phase, while the amplifier gain and coupler loss are chosen for a loop-gain amplitude of slightly greater than unity. To measure the loop gain and phase, the switches in Figure 6.12 are set to open circuit the feedback loop. The measurements in Figure 6.13 show that a zero phase crossing occurs around 6 GHz, and that there is loop

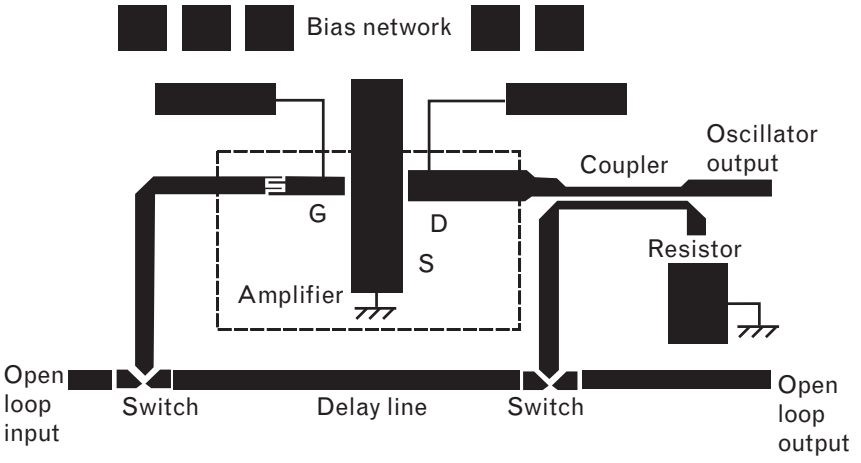


FIGURE 6.12 *Schematic diagram of an FET with feedback that can be switched between open-loop (amplifier) and closed-loop (oscillator) conditions. (From: [4]. © 1992 IEEE. Used with permission.)*

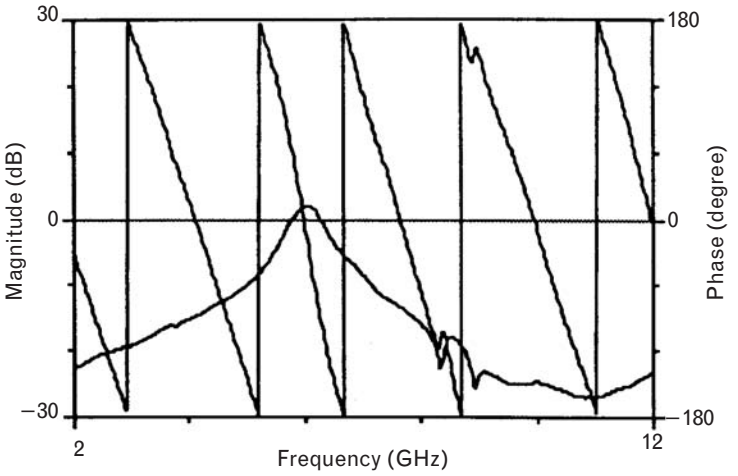


FIGURE 6.13
Measurement of open-loop gain and phase for the circuit of Figure 6.12. (From: [4]. © 1992 IEEE. Used with permission.)

gain greater than one at that frequency.[4] The extra phase introduced by the feedback delay line needs to be accounted for in determining the zero crossing point. When the loop was closed, an oscillation frequency of 6.05 GHz was measured, within 1% of the estimated value. The spectrum of the output signal was observed to be free of any spurious effects and out-of-band oscillations. This was ensured since at frequencies away from 6 GHz, the gain was designed to be always less than one so that the possibility for oscillations there was suppressed.

The design of oscillators using open-loop synthesis is detailed in a number of application notes from the test equipment manufacturers and is their preferred methodology for design (because they can sell more network analyzers!). The technique, being nonautonomous, allows the system to be driven and simulated by an external signal at a known frequency, rather than simulated free-running as the case must be when the input port is lost by closing the loop. Then, because tuning is performed open-loop, cause-and-effect can be more clearly understood. The open-loop analysis enables any unusual responses due to out-of-band resonances to be examined, and indicates the possibility of other potential oscillation frequencies at which the Barkhausen criterion could be met. In theory, it also enables the device load line to be designed to avoid saturation and regions of forward conduction (i.e., voltage-limited regions) that can degrade the phase noise.

6.1.2 One-port oscillator design approach

6.1.2.1 A series resonant circuit as an oscillator

We have illustrated that the consequence of closing the loop of a system designed to meet the Barkhausen criterion creates a negative resistance

4. Here, the authors have approximated the open-loop gain by s_{21} instead of the parameter G, which would be more accurate.

seen looking into the output port. This is a quite general result and is true of any oscillating circuit no matter where the loop is cut to create an "output" port. We will use this observation in the simple circuit of Figure 6.14 to derive some fundamental expressions that are helpful in analyzing the behavior of an oscillator.

Assume that the excitation voltage in the figure is the bias voltage, which is switched on at $t = 0$. We can apply Kirchhoff's voltage law to the circuit to obtain

$$e_{IN} = -|R_D|\, i + \frac{1}{C}\int_0^t i\,dt + L\frac{di}{dt} + R_L i \tag{6.6}$$

or if we use the Laplace transform[5] of the circuit and consider e_{IN} to be a step voltage, then

$$\frac{e_{IN}}{s} = \left(R_L - |R_D|\right) i + \frac{i}{sC} + sLi \tag{6.7}$$

or solving for the current

$$i = \frac{e_{IN}}{s^2 L + \left(R_L - |R_D|\right)s + \frac{1}{C}} \tag{6.8}$$

The output voltage taken across R_L may then be written as iR_L or

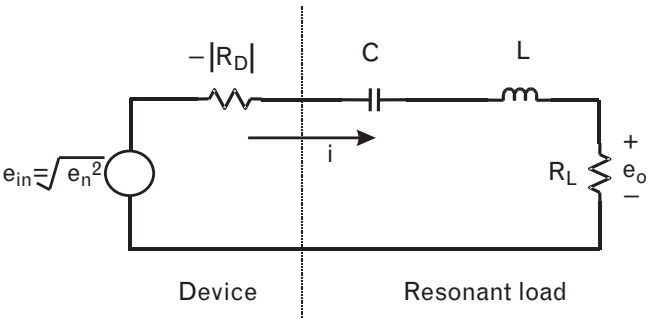


FIGURE 6.14
Series oscillator circuit, showing the device resistance, resonant circuit, and load resistor.

5. The lowercase letter *s* without any subscript refers to the complex variable $s = j\omega$, not to an *S*-parameter.

$$\begin{aligned} e_O &= e_{IN}\,\frac{R_L}{L}\,\frac{1}{s^2 + \dfrac{\left(R_L - |R_D|\right)}{L}s + \dfrac{1}{LC}} \\ &= \left(e_{IN}\,\frac{R_L}{L}\right)\frac{1}{s^2 + 2\zeta\omega_0 s + \omega_0{}^2} \end{aligned} \tag{6.9}$$

where

$$\begin{aligned} \omega_0{}^2 &= \frac{1}{LC} \\ \zeta &= \frac{\left(R_L - |R_D|\right)}{2\omega_0 L} = \frac{1 - \left|\dfrac{R_D}{R_L}\right|}{2Q} \\ Q &= \frac{\omega_0 L}{R_L} \end{aligned} \tag{6.10}$$

The roots of (6.9) are given by

$$s = \omega_0\left(\zeta \pm j\sqrt{1-\zeta^2}\right) \tag{6.11}$$

so the inverse Laplace transform of (6.9) is given by

$$e_o = \left(e_{IN}\,\frac{R_L}{L}\,\frac{1}{\omega_0\sqrt{1-\zeta^2}}\right)e^{-\zeta\omega_0 t}\,\sin\left(\omega_0 t\sqrt{1-\zeta^2}\right) \tag{6.12}$$

Equation (6.12) is plotted in Figure 6.15 for both positive and negative values of ζ.

The time-domain response given by (6.12) is a sinusoid of frequency $\omega_0\sqrt{1-\zeta^2}$ multiplied by an envelope of value $e^{-\zeta\omega 0t}$. The envelope can be written as $e^{-t/\tau}$, where τ is the time for oscillations to decay to $1/e$ of their initial value. The τ is equal to $1/\omega_0\zeta$ so is directly proportional to Q. For positive values of ζ, the envelope decays over time and there is no steady-state oscillation. However, for negative values of ζ, oscillation grows exponentially because the envelope increases with time.

From this, we deduce (initially, at least) that the conditions for startup of oscillation are

$$\zeta < 0 \rightarrow |R_D| > R_L \tag{6.13}$$

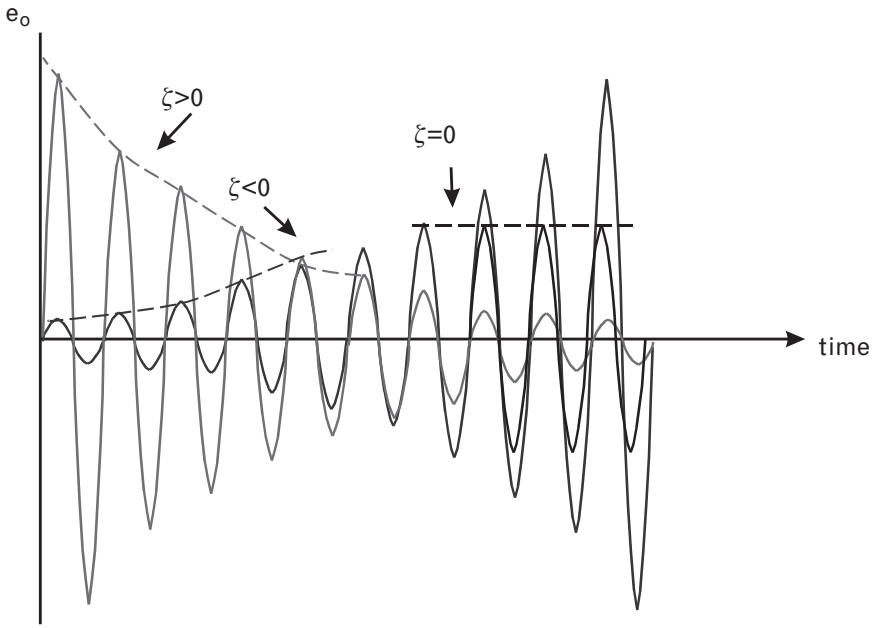


FIGURE 6.15
Output voltage of the series RLC circuit to a step response at the input.

where we have used the definition of ζ in (6.10), and of course, the device resistance is assumed negative throughout the analysis. Eventually, the system must limit to a point where ζ equals zero, so that the envelope equals unity and the amplitude of the sinusoid in (6.12) is constant. Then, the oscillation frequency given by (6.12) is just ω_0. At this frequency, $\omega_0 L = 1/\omega_0 C$ from the first equation in (6.10), so the net reactance around the loop equals zero. Although we have not explicitly shown any reactance associated with the device in Figure 6.14 for simplicity, it is straightforward to include a reactive component jX_D in series with $-|R_D|$, and the result just derived implies this will net out the reactance of the resonator at the frequency of oscillation. Thus, at steady state, the following conditions apply:

$$\begin{aligned} &\zeta = 0 \rightarrow |R_D| = R_L \rightarrow R_D = -R_L \\ &\omega = \omega_0 \rightarrow X_D + X_L = 0 \rightarrow X_D = -X_L \end{aligned} \tag{6.14}$$

These well-known results form the basis of one-port oscillator theory. The magnitude of the device resistance will equal the load resistance at steady state, where from the assumptions made in (6.6) the device resistance is a negative quantity. The load reactance is equal and opposite to the device reactance because at resonance there is no net reactance in the loop. These results suggest an alternative mechanism for oscillator design: configure a device to have a negative resistance looking into its output port, and terminate that port with a load resistance equal and opposite to its

device resistance and a load reactance equal and opposite to the device reactance. This is true, but it is only part of the story.

6.1.2.2 The negative resistance oscillator

In the above analysis, we have purposely used $-|R_D|$, for the device resistance to highlight that it is negative in an oscillator. In the following, we now assume that for an oscillator the device resistance R_D is a negative number itself and omit the explicit magnitude bars and negative sign for more generality. Equations (6.13) and (6.14) together imply that as the signal level builds up in an oscillator during startup, the device resistance changes. In the example of the series circuit shown, the negative device resistance must start out with a negative value that is larger (more negative) than the load resistor. As oscillations increase, it must become less negative until it is equal and opposite to the load resistance. This principally occurs because as the drive at the input of the transistor increases as the oscillations grow, the device transconductance is reduced as a consequence. Such a circuit is characteristic of a closed-loop system with (positive) series feedback.

If the amplitude of the signal swing is characterized by its amplitude A, the impedance looking into the output port of an oscillator is a strong function of that amplitude. In the case of Figure 6.14, the amplitude is taken as the current that flows around the loop. At the frequency of oscillation ω_0, we define

$$Z(A)\big|_{\omega_0} \triangleq -Z_D \tag{6.15}$$

where $Z_D = R_D + jX_D$ is the measured or simulated device impedance seen looking into the output of the active device. The terminals of interest at which we choose to split the so-called device from the rest of the circuit can be chosen fairly arbitrarily in Figure 6.14. However, it is best to associate the bias network and any terminating impedances with the device, and to associate the resonant circuit, or main frequency determining components of the oscillator with the "load" shown as R_L in the figure. It is clear that Figure 6.14 is overly simplistic in that the resonant circuit is represented by a simple series L-C circuit, and the resonator losses and actual load impedance are lumped into a single load resistor R_L. Usually the device itself will also include a reactive component jX_D that will probably vary with drive level as well. This concept of splitting an oscillator into a device or active part, and a load or resonant part (or into an "osci" and a "llator" as one author has suggested [5]) proves useful.

Thus, (6.15) becomes

$$Z(A)\big|_{\omega_0} = R(A) + jX(A) = -R_D - jX_D \tag{6.16}$$

measured at the oscillation frequency. If the total "load" on the device is then characterized as a function of frequency (since it is a linear element, and presumably frequency sensitive since it contains the resonator), we may rewrite the conditions for steady state oscillation (6.14) as

$$\begin{aligned} R(A) &= R_L(\omega) \\ X(A) &= X_L(\omega) \end{aligned} \tag{6.17}$$

These are defining equations and a necessary condition for oscillation. They indicate that steady-state oscillation is possible at that impedance point at signal level A and oscillation frequency ω_0 for which the device resistance and impedance are equal and opposite to the load impedance. The device impedance is a strong function of signal level, but also of frequency since a transistor will vary in impedance with frequency as well. But compared to the load circuit, which contains the resonator and is thus a strong function of frequency about the oscillation point, we typically neglect the frequency dependence of the device in the vicinity of oscillation and characterize it, as in (6.16), as more strongly variant with drive level.

This is illustrated in Figure 6.16 where we show $Z(A)$, the negative of the device impedance for the base of a common-collector bipolar transistor, with the emitter terminated in a capacitance. As the drive is increased from small-signal conditions to large-signal, by increasing the power incident on the base, $Z(A)$ moves along a curve of fairly constant reactance, but its resistance decreases from around 30Ω to 10Ω. Thus the device resistance itself becomes less negative, from –30Ω to –10Ω , as the signal grows. Terminating the base with a load resistor of 10Ω will (typically) result in steady-state oscillations, since (6.13) is satisfied for small-signal levels (i.e.,

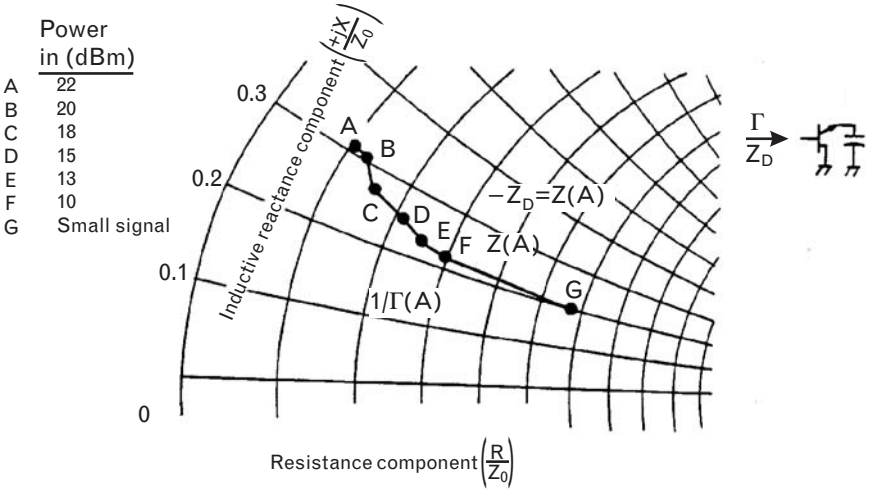


FIGURE 6.16
The negative of the device impedance looking into the base of a common-collector transistor with emitter terminated by a capacitor.

ζ is negative so oscillations can grow with an exponentially increasing envelope). At large signal levels (6.14) is satisfied and ζ is zero, so the envelope of the sinusoid is constant. On the other hand, a load resistor of 50Ω at the base results in a positive ζ, so according to (6.12), any oscillations will be damped by an envelope that decays exponentially over time.

The above procedure for terminating a negative impedance device to result in oscillations is so widespread that it has resulted in a rule of thumb that has become part of RF and microwave folklore: Terminate a negative resistance device with a load resistance that is approximately one-half to one-third of the small-signal value of the device impedance. In the above example, this translates to terminating the base with a load resistor of between 10Ω and 15Ω. This, according to the rule of thumb, means that at small signals the device resistance will be sufficiently more negative than the load is positive, so that ζ is negative and oscillations will grow. Unfortunately, like most rules of thumb, it is true, some of the time. In this case, it is wrong for half of the time.

In fact, this rule of thumb was derived for IMPATT diode oscillators many years ago, prior to the appearance of more modern devices that can be used for high-frequency oscillators. IMPATT diodes can be modeled fairly precisely by a fifth-order nonlinearity, and as a function of current drive, the optimal large signal operating point has a negative resistance around one-third of the small-signal value, but this is not at all the case for other devices.

However, there is a more serious problem with this rule, and that is that it is only applicable for series circuits of the form shown in Figure 6.14. Consider instead the parallel circuit of the form shown in Figure 6.17.

This circuit is the exact dual of the circuit of Figure 6.14. Here, instead of characterizing the output voltage as a function of loop current, the output current can be derived as a function of node voltage. Previously, the voltage was expressed as the product of current and impedance; now, the current is expressed as the product of voltage and admittance. Such a circuit is characteristic of a closed-loop system with (positive) shunt feedback. As before, oscillations will grow when $\zeta < 0$ and stabilize with constant envelope once $\zeta = 0$. But the defining equations are now the dual of those in (6.13) and (6.14):

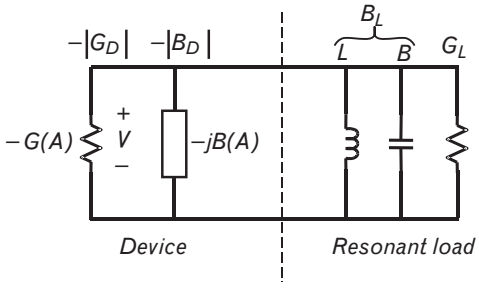


FIGURE 6.17
Shunt type oscillator model with a parallel resonant circuit.

$$\zeta < 0 \rightarrow |G_D| > G_L \tag{6.18}$$

and

$$\begin{aligned} \zeta = 0 &\rightarrow |G_D| = G_L \rightarrow G_D = -G_L \\ \omega = \omega_0 &\rightarrow B_D + B_L = 0 \rightarrow B_D = -B_L \end{aligned} \tag{6.19}$$

Such conditions are in fact the exact opposite of those that would be predicted by applying the rule of thumb, which is valid only for the series circuit. Consider again the impedance variation of Figure 6.16. The device conductance with drive actually becomes more negative, decreasing from –0.0333 siemens (–30Ω) at small-signal levels, to –0.10 siemens (–10Ω) at large-signal levels. In this case, a 10-Ω (0.1-siemens) load resistance would satisfy (6.19) for steady-state oscillation but would violate (6.18) at small-signal levels, so that oscillations could never increase initially.

Clearly, we need to be able to discern whether the device behaves as a negative resistance device in which the negative resistance becomes less negative with increasing signal (current), or as a negative conductance device in which the negative conductance becomes less negative with increasing signal (voltage). Figure 6.18 summarizes the two dual one-port oscillators we have discussed so far:

1. A series feedback oscillator, which will not oscillate with large load resistances because the loop resistance ($R_L + R_D$) is always so positive that $\zeta > 0$ and oscillations can never build up;
2. A shunt feedback oscillator, which will not oscillate with large load conductance because the loop conductance ($G_L + G_D$) is always so positive that $\zeta > 0$ and oscillations can never build up.

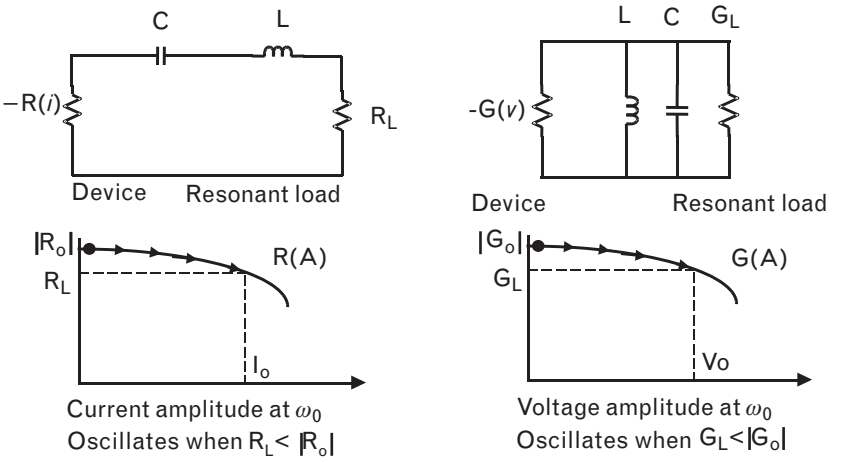


FIGURE 6.18
Behavior of the load resistance and load conductance of the two types of oscillators, the series and shunt configurations. R_O and G_O are negative and are the small-signal device resistance and conductance, respectively.

We also need to consider how we might create either a negative resistance or a negative conductance in the first place. We will see later in this chapter how to do this through simulation, but we need to first complete our analysis of startup conditions in an oscillator.

6.1.2.3 Oscillator startup—more detailed considerations

The denominator of the closed-loop gain expression (6.2) is known as the characteristic equation for an oscillator. When it equals zero at the expected oscillation frequency and amplitude, the Barkhausen criterion

$$\begin{aligned} \Re e\left(AL\left(j\omega_0\right)H\left(j\omega_0\right)\right) &= 1 \\ \Im m\left(AL\left(j\omega_0\right)H\left(j\omega_0\right)\right) &= 0 \end{aligned} \tag{6.20}$$

results. However, this is not on its own a good indicator of instability because it does *not* ensure the presence of right-half plane poles at startup.

For a feedback system to oscillate, its closed-loop gain must have a pair of complex-conjugate poles in the right-half plane. If the poles are given by $p_{1,2} = \alpha \pm j\beta$, then a necessary condition for oscillation to start is that $\alpha > 0$ so that the envelope

$$x(t) = Ke^{\alpha t}\cos(\beta t) \tag{6.21}$$

is exponentially growing. The location of these poles will be a function of the gain, and a root-locus plot shows their location as the gain increases. Ideally, the poles will move towards the left-half plane as the gain reduces, and cross the imaginary axis at some frequency $\beta = j\omega_0$, corresponding to the oscillation frequency. At this point, $\alpha = 0$ and the envelope given by (6.21) is of constant amplitude.

The existence of right-half plane poles is a *necessary and sufficient* condition for oscillations to grow. As a rule of thumb, (6.20) is a sufficient condition if it holds at only one frequency ω_0 [6]. Similarly, the conditions given earlier for the negative-resistance oscillator to start up (or its admittance dual for negative-conductance oscillators)

$$\begin{aligned} R\left(A,\omega_0\right) &> R_L\left(\omega_0\right) \\ X\left(A,\omega_0\right) &= X_L\left(\omega_0\right) \end{aligned} \tag{6.22}$$

are necessary but not sufficient. Equation (6.22) is generally a sufficient condition if the reactance condition is satisfied at only one frequency rather than at multiple frequencies. This requirement can be satisfied if the total reactance—the sum of the device and the load reactance—is monotonic with frequency. This turns out to be satisfied when we load a negative-

resistance device with a series resonant circuit or a negative-conductance device with a parallel resonant circuit.

Equation (6.22) fails to guarantee startup behavior because it applies only at a single frequency, and then for small-signal levels. Such single-point analysis fails to consider the frequency and drive behavior of the device as oscillations build up. Because the root-locus plot shows the location of the device poles with *both* frequency and gain, it can provide more accurate information on oscillation than our earlier simplistic approaches. One objective of oscillator design using the large-signal simulation techniques outlined in Chapter 4 should therefore be to ensure that the embedding circuit produces just one pair of right-half plane poles in the closed-loop gain function, lying close to the $j\omega$-axis.

Unfortunately, the root-locus plot is usually neither convenient nor simple to obtain—let alone remembered by most RF engineers! Therefore, a Nyquist plot of the small-signal open-loop gain is more convenient. The Nyquist plot is just a plot of $AL(j\omega)H(j\omega)$—the same gain function used for the more common Bode plot[6]—drawn on a polar plot as a function of frequency. It can confirm the existence of right-half plane poles, since the number of net clockwise encirclements of the point (1,0) when the open-loop gain is plotted for $-\infty < j\omega < +\infty$ on the polar plot equals the number of such poles.[7]

The Nyquist stability criterion can also be checked using the Bode plot of the small-signal gain. When the open-loop gain is greater than one at the zero-phase crossing, negative phase slope corresponds to clockwise rotation about the point (1,0) on the Nyquist plot. Positive phase slope corresponds to counterclockwise rotation. This is illustrated in Figure 6.19. When the gain is less than one, those directions are reversed.

An example of a Nyquist plot with gain less than one is given for the circuits in Chapter 4, although in that example it was the product of the device and circuit reflection coefficients that was plotted to predict instability, rather than open-loop gain, since the oscillator characteristic equation there was framed in terms of device and load reflection coefficient. A net clockwise encirclement of (1,0) can be ensured when the small-signal open-loop gain is greater than 1 and the phase slope is negative at a single zero-phase crossing. This is then sufficient to guarantee oscillator startup.

6. The Bode plot is simply the rectangular plot of gain and phase as a function of frequency on the x-axis.

7. In Chapter 1, the Nyquist stability criterion was expressed in terms of encirclement of the point (–1,0), because there the negative of the open-loop gain was plotted. Functionally, this is identical. It is more common in the stability analysis of amplifiers, where negative feedback, not positive, is desirable. Changing the summing node of the oscillator model in Figure 6.1 to a subtraction node transforms the model to such an amplifier with negative feedback. With oscillators, when the open loop gain equals +1, the denominator of the closed-loop gain equation becomes zero and right-half plane poles may exist, but not necessarily at the required value of frequency and gain to ensure sustainable oscillation.

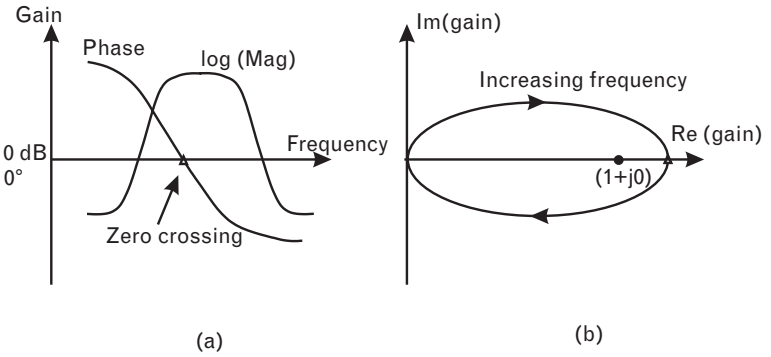


FIGURE 6.19
Comparison of (a) the Bode plot and (b) the Nyquist diagram of the oscillator characteristic equation, showing the open-loop gain for a typical oscillator with a single zero-crossing of the phase, where it has negative slope and excess gain.

Properly designed oscillators will use series resonant terminations on negative-resistance devices and parallel resonant terminations on negative-conductance devices. For such oscillators, the open-loop gain generally has only one zero-phase crossing and it must have negative slope (as in Figure 6.19) with magnitude greater than one. Under such conditions, (6.13) or (6.18), or (6.20), are sufficient alone to ensure startup of oscillation.

As an example, consider the oscillator analyzed by Nguyen [6]. This is a Pierce oscillator whose schematic is shown in Figure 6.20. By replacing the bipolar transistor with its small-signal equivalent model and breaking the loop, one can show that the open-loop gain for this circuit can be written in the form

$$AL(s)H(s) = -A_0 \frac{s^2 LC + s\dfrac{L}{R} + 1}{a_3 s^3 + a_2 s^2 + a_1 s + a_0} = T(s) \tag{6.23}$$

where the coefficients in the denominator are functions of the external circuit and where $A_0 = g_m R_1$. The poles of the circuit can be found by solving

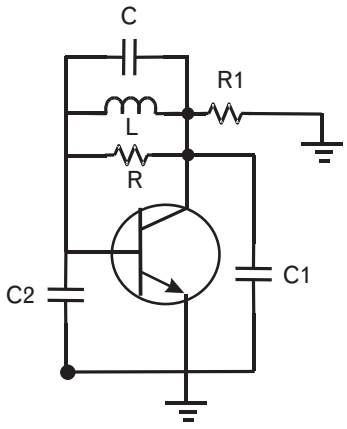


FIGURE 6.20
Circuit schematic for the Pierce oscillator. (From: [6]. © 1992 IEEE. Used with permission.)

for the values of the complex frequency s for which the characteristic equation $1 - AL(s)H(s) = 0$, since the closed-loop gain is given by $AL(s)H(s) / [1 - AL(s)H(s)]$. Plotting these values of s as a function of the gain A gives the root-locus plot.

Figure 6.21(a) shows the root-locus plot as a function of the transconductance g_m. Between the two values of g_{m1} and g_{m2} the complex-conjugate poles enter the right-half plane. Here, the circuit is unstable and oscillations can grow. If the loop gain is set too large, however, the complex poles reenter the left-half plane and the circuit becomes stable again, even though (6.20) is satisfied. We can see this on the Bode plot in Figure 6.21(b) drawn for $g_m > g_{m2}$. This plots the magnitude and phase of the open-loop gain $T(s)$ in (6.23) as a function of frequency. Even though the gain and phase conditions for steady-state oscillation are indeed satisfied at two frequencies, 1,020 and 1,180 MHz, the root-locus plot shows there are no right-half plane poles for $g_m > g_{m2}$. Equation (6.20) is thus a necessary condition for oscillation, but it is not sufficient. The presence of multiple

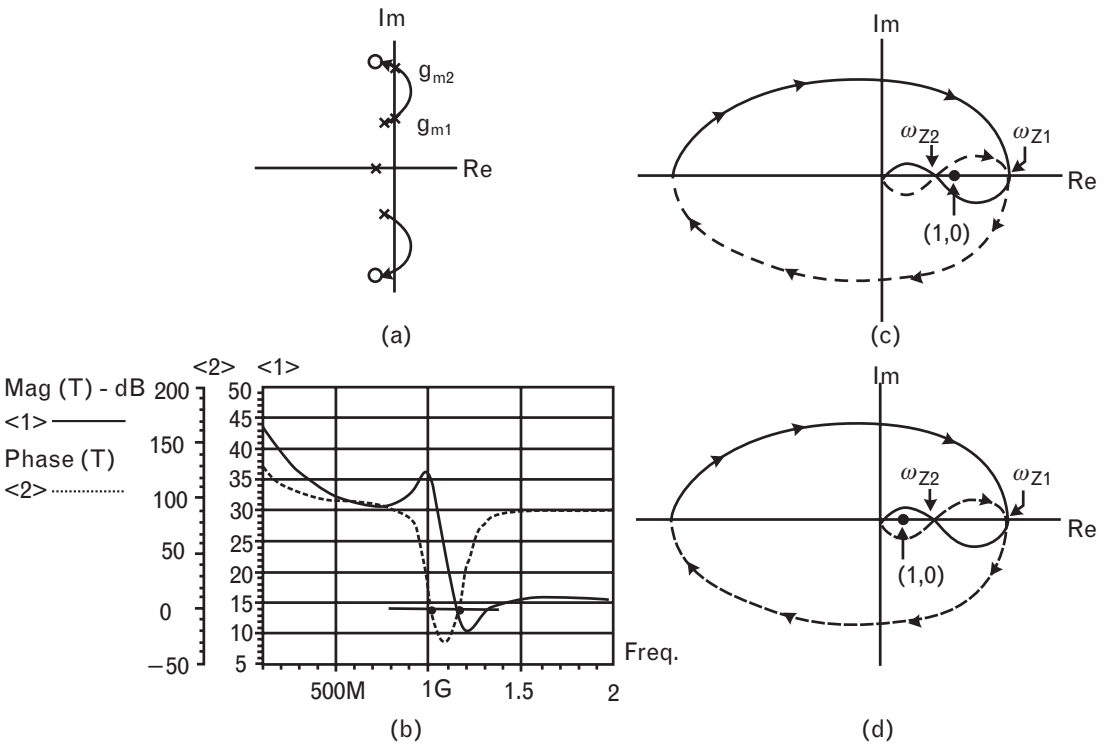


FIGURE 6.21 *Analytical plots for the Pierce oscillator circuit of Figure 6.20 and its loop gain $T(s) = AL(s)H(s)$. (a) Root locus. (b) Bode plot where $g_m > g_{m2}$. (c) Nyquist diagram, for gain $g_{m1} < g_m < g_{m2}$ that has two complex right-half plane pole pairs. (d) Nyquist diagram, for stable gain $g_m > g_{m2}$. The dashed-line indicates the negative frequencies. (From: [6]. © 1992 IEEE. Used with permission.)*

zero-phase crossings should be used as an indicator that the poles may not remain in the right-half plane, and oscillation will not occur.

The Nyquist plot of the open-loop gain reveals more. It is shown in Figure 6.21(c) for a value of $g_{m1} < g_m < g_{m2}$ where the root locus indicates that right-half plane poles do exist. The point (1,0) has two clockwise encirclements, indicating two right-half plane poles (one complex conjugate pair). On the other hand, the Nyquist plot in Figure 6.21(d) for $g_m > g_{m2}$ has no clockwise encirclements of (1,0) and confirms that the circuit is stable for this value of g_m. In summary, the Nyquist plot of the oscillator open-loop gain can confirm the existence of right-half plane poles, which is a sufficient condition to ensure startup of oscillation. Excess small-signal gain at the zero-phase crossing is necessary but not sufficient to guarantee oscillations.

Randall and Hock [3] have extended startup analysis to linear two-port *S*-parameters, so that oscillator stability can be considered at both small- and large-signal levels using the appropriate CAD simulation. This avoids the need to characterize the active device as above in terms of its transfer function for the root-locus plot and allows simple measured quantities to be used instead.

The *S*-parameters of the oscillator circuit recast as an open-loop system are required, with the active device embedded within the resonator and load circuits. A convenient point in the oscillator topology first needs to be established to break the feedback loop at some point to calculate the two-port *S*-parameters looking into the break. Ideally, the circuit topology will be recast choosing an ac ground and break point in such a way that the oscillator in open–loop most resembles a cascaded amplifier and a tuned circuit, and to minimize the reverse transfer characteristic s_{12}. The loop can then be simply closed for oscillations by reconnecting the output node directly to the input. Although convenient for understanding, a well-chosen split point is not essential for accurate modeling.

If an oscillator is recast as a feedback system in this way, and a break point is chosen, we saw earlier in this chapter that the open-loop gain can be calculated as

$$G = \frac{s_{21} - s_{12}}{\left(1 - s_{11}s_{22} + s_{12}s_{21} - 2s_{12}\right)} \tag{6.24}$$

where the *S*-parameters are those of the entire open-loop oscillator circuit. Although it might be expected that the open-loop gain should be just s_{21} alone, we saw in Section 6.1.1.2 that this does not account for the effects of impedance mismatch at the connection point, or reverse gain, when the loop is closed.

Whenever the open-loop gain G is equal to one in amplitude and zero in phase, the closed-loop transfer function (6.2) has a pole. The condition $G = 1$ is satisfied whenever

$$1-\left(s_{12}+s_{21}-s_{12}s_{21}+s_{11}s_{22}\right)=0 \tag{6.25}$$

This condition can also be derived independently by writing the ratio of the output to input current of Figure 6.22 in terms of its *S*-parameters, and solving for the roots when the input current is set to zero. It is an alternate expression for the characteristic equation of an oscillator.

G can now be used in Nyquist and Bode plots in a similar fashion as before to predict circuit stability. Thus, when *G* is plotted on a Nyquist plot for increasing frequency for $-\infty < f < \infty$, the polar plot must make at least one net clockwise encirclement of the point (1,0) as the frequency is increased in order for the circuit to be unstable.

If the phase shift of the device itself near the oscillation frequency is ϕ, then this is compensated for by a phase shift $-\phi$ in the resonator and load circuits, so the closed-loop phase sums to zero. The oscillation frequency then shifts to a new frequency of zero-phase shift

$$\omega_0 = \omega_{LC}\left(\frac{\tan(\phi)}{2Q_0}+\sqrt{1+\frac{\tan^2(\phi)}{4Q_0^{\,2}}}\right) \tag{6.26}$$

where $\omega_{LC} = 1/\sqrt{LC}$ is the resonant, or zero-phase frequency of the resonator and load circuits, and Q_0 is the Q of the resonator with load. The loaded Q of the oscillator can be computed from the phase slope at the zero-phase point using

$$\begin{aligned} Q &= -\frac{1}{2}\omega_0\frac{\partial\varphi}{\partial\omega} = \frac{1}{2}\omega_0 t_g \\ &\approx \frac{\omega_0}{\omega_{LC}}Q_0\cos^2\phi \end{aligned} \tag{6.27}$$

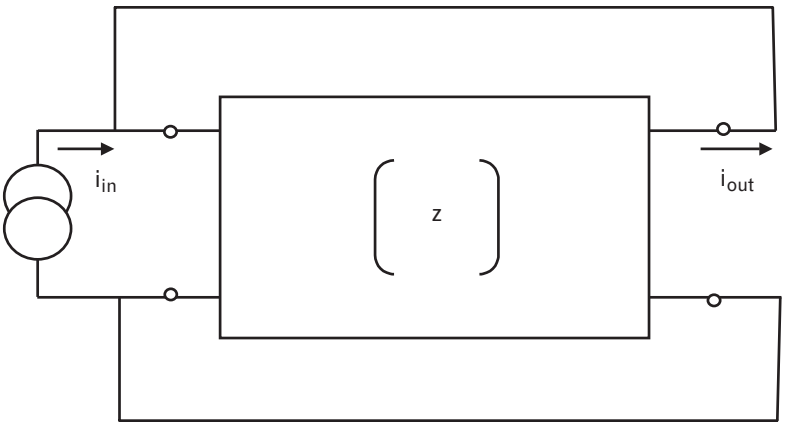


FIGURE 6.22
A closed-loop feedback system to model an oscillator and driving source. (From: [3]. © 2001 IEEE. Used with permission.)

where t_g is the group delay of the open-loop circuit. The loaded Q is therefore lower than the Q of the resonator with load by a factor that depends on the phase offset ϕ through the device itself.

Example of a Nyquist plot

Consider again the oscillator circuit of Figure 6.9, whose open-loop gain G and phase are plotted in the Bode plot of Figure 6.7(a). When the same function is plotted on a polar plot from $0 < f < \infty$, the Nyquist plot of Figure 6.23(a) results. (Negative frequencies are required to complete the plot, but they will be the mirror image of positive frequencies and may be simply visualized.)

By design, the Nyquist plot passes through the point (1,0) at 1,000 MHz since this is the design condition for steady-state oscillation, when large-signal S-parameters are used to account for the reduced gain of the device when limiting. However, when the device small-signal S-parameters are used instead to reflect the state of the system at startup, the Nyquist plot of Figure 6.23(b) results. At zero phase, which now occurs slightly above 1 GHz, the magnitude of the gain G is 1.0045 and crosses to the right of the point (1,0). The phase slope is negative at this point. The trace moves clockwise as the frequency increases from 0 to ∞, and since negative frequencies plot as the mirror image of those shown, the point (1,0) is encircled twice by the small-signal open-loop gain, a sufficient condition to ensure startup of this oscillator.

Interestingly enough, the value of G at 1,000 MHz is not relevant to startup, since in this case it is the value of G at a higher frequency (1,008

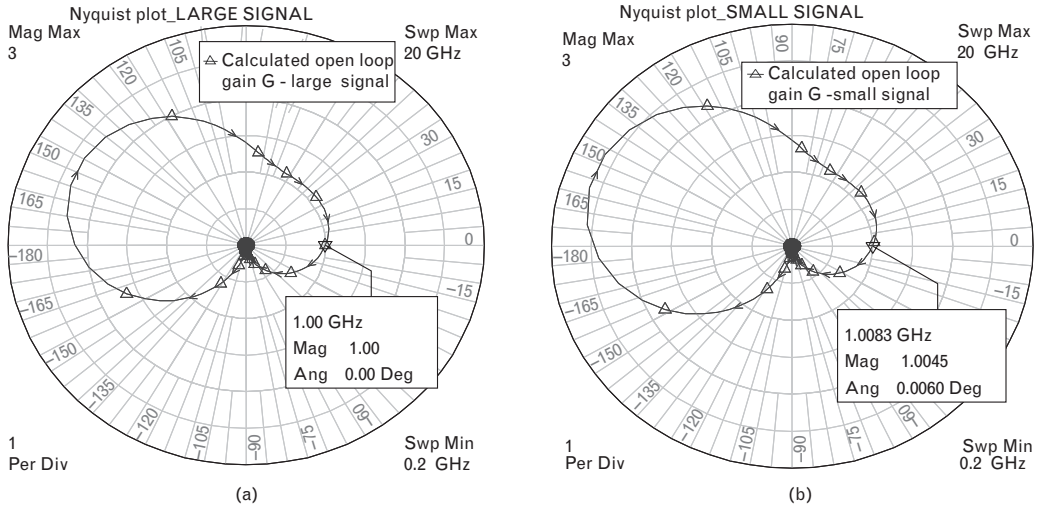


FIGURE 6.23 *Nyquist plot for the circuit of Figure 6.9 (a) using large-signal S-parameters at the steady-state conditions, and (b) using small-signal S-parameters. Only positive frequencies are plotted. Negative frequencies will plot as the mirror image.*

MHz) that determines whether or not the point (1,0) will be encircled, or not at all. We should also note that G is zero at dc only because a highpass L-C circuit was used in cascade with the (open-loop) device. The shunt inductance of this L-C resonant circuit effectively removes any low-frequency open-loop gain. This is a contributing factor to the encirclement of (1,0). Were a lowpass resonant circuit used instead (e.g., a shunt capacitor–series inductor combination), the gain at dc would be high and the point (1,0) would not be encircled at all. This plot is a simple and effective way to check the effect of different feedback topologies on the loop gain and startup conditions, across all frequencies.

6.1.2.4 Characterization of the oscillator negative impedance

In the late 1960s, Kurokawa [7] performed extensive analysis on the locking and noise properties of an oscillator, through an analysis of the device impedance and its interaction with the load. As discussed earlier and as embodied in (6.17) (or its dual equation), a necessary condition for steady-state oscillation is that at the frequency of oscillation, the amplitude of the signal swing will be such that the device impedance adjusts to be equal and opposite to the load impedance. This ensures that $\zeta = 0$ and that oscillations have reached a constant amplitude.

This condition can be shown graphically. If we plot both the resonant load impedance as a function of frequency, and the negative of the device impedance (measured at or near the oscillation frequency) as a function of signal amplitude, the point of intersection will give the impedance at which the load impedance is equal and opposite to the device impedance. The negative of the device impedance is, of course, just $Z(A)$ from (6.15). The plot can be drawn either on Cartesian coordinates with real and imaginary parts of the impedance as the axes, as shown in Figure 6.24, or on the Smith chart. The plot of the load impedance $Z_L(f)$ as a function of frequency is made with an arrow in the direction of increasing frequency to

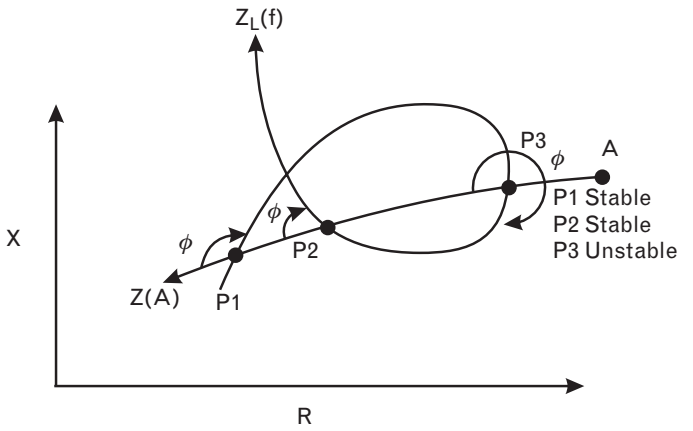


FIGURE 6.24
The device and load line locus. The device line Z(A) is a function of signal amplitude A at the oscillation frequency; the load line $Z_L(f)$ is a function of frequency f. Both are vectors in the direction of increasing amplitude or frequency.

indicate the vector, and the plot of $Z(A)$ is made with an arrow in the direction of increasing signal level A.

Kurokawa also proved (after considerable algebra) two other important conditions:

1. If the crossing angle is defined as the vector angle between $Z(A)$ and $Z_L(f)$, then only intersections with angles between zero and 180° are stable oscillating points. Angles greater than 180° will not result in sustainable oscillation. The angle is measured from $Z(A)$ to $Z_L(f)$. A *stable* oscillation is one for which any perturbation in amplitude A will decay with time.
2. A crossing angle of 90°corresponds to minimum phase noise; a crossing angle of zero or 180° corresponds to the noisiest conditions.

In Figure 6.24, points P1 and P2 are stable operating points; point P3 is unstable because the crossing angle, from $Z(A)$ to $Z_L(f)$ measured in the direction of increasing A and f, is close to 270°. These conditions cannot only help us to explain effects such as mode hopping and hysteresis in oscillators, but also to design the best load to terminate a device to achieve minimum oscillator phase noise.

Figure 6.25 illustrates this. Suppose the device line $Z(A)$ shifts upward due to some environmental effect. This effect could be temperature, or perhaps a bias change on the device to tune the frequency. As a result of changing the tuning parameter from T1 to T3, the device reactance changes and the points of intersection with the line $Z_L(f)$ shift. Thus at T1, there is a single steady-state oscillation frequency f_0, at a relatively high signal level (because A is high). This is a stable operating point and because the crossing angle is close to 90°, there is relatively low phase noise at this point. However, as the tuning curve is increased, to a position between T1 and T2, multiple intersections begin to occur. The frequency will remain along the same section of load line. For instance, at T2, the oscillations will

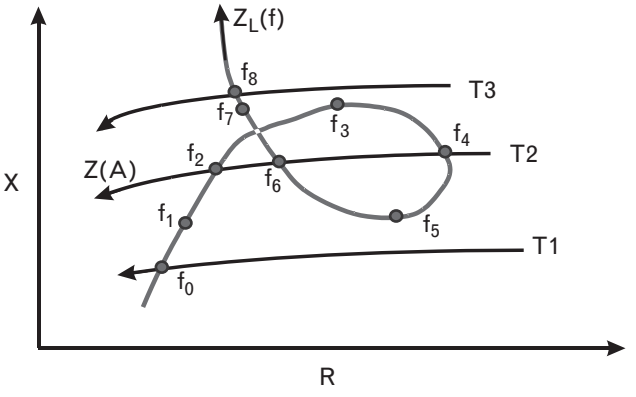


FIGURE 6.25
A VCO, in which the device line is tuned with bias or temperature, shifting the oscillation point.

remain at f_2 because this is a stable operating point and there is no perturbation that can push the frequency to f_6 or f_4. Although f_6 is a valid (stable) operating point, the oscillator will remain oscillating at f_2 because it has approached this frequency by tuning from f_1 and remains stable. As the tuning continues to increase, the frequency eventually increases to f_3. At this tuning level, the point of intersection becomes a line of intersection, and any noise in the tuning level violently shifts the frequency about f_3. This point, with a crossing angle of 180°, can be seen intuitively to be very noisy. At higher tuning levels, the frequency makes a hop to a frequency f_7, which is once again a stable oscillating point. As the tuning level continues to increase, the frequency once again rises smoothly through f_8. The oscillator has made a mode hop in jumping from f_3 to f_7, and the frequencies between them cannot be reached through this tuning route.

Unfortunately the tuning history of this oscillator is not as simple as a single mode-hop. As the tuning level is retraced, from T3 to a level back towards T2, the intersection point retraces a different part of the $Z_L(f)$ curve. Because f_6 is a stable operating point, the frequency retraces from f_8 through f_6 until f_5 is reached, where once again the "point" of intersection becomes a line of intersection and the frequency of oscillation is very noisy. As the tuning level is further reduced, a mode-hop occurs down to a frequency f_1, so that frequencies between f_5 and f_1 cannot now be reached by this tuning route. As the tuning is further reduced, we retrace the same tuning curve as before down to f_0.

The tuning curve for this oscillator is shown in Figure 6.26. It is apparent that not only does this oscillator suffer from two mode-hops, but that it also has hysteresis. This unfortunate situation, which can be observed in many real oscillators, arises because the load on the oscillator offers a load impedance to the device that sustains oscillation at more than one frequency. If, in fact, the load line $Z_L(f)$ were a straight vertical line of constant real part R and variable reactance X with frequency, there would only

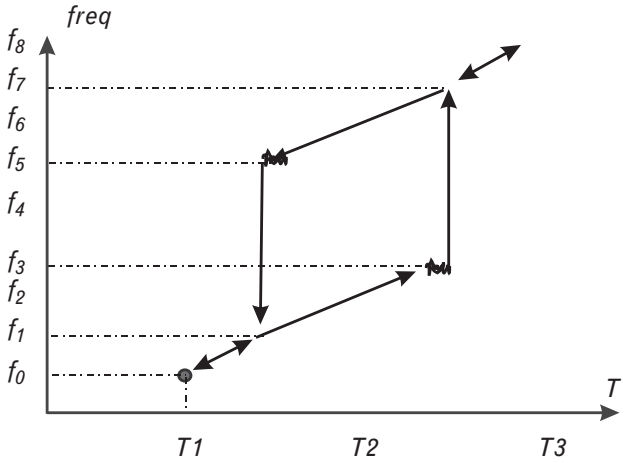


FIGURE 6.26
The tuning curve of the oscillator represented by the device and load line locus of Figure 6.25.

be a single possible oscillation point at each tuning level. Such a load could be synthesized with a series RLC circuit: the real part is just the series resistance, which is constant, and the imaginary part (or reactance) will be given by $\omega L - 1/\omega C$, which increases monotonically with frequency $f = \omega/2\pi$.

In fact, the load impedance shown in Figure 6.25 is not at all a complicated circuit. Consider the circuit of Figure 6.27, which is a parallel RLC circuit and a small inductance in series, as might occur with a bond wire. On the Smith chart, a parallel RLC circuit on its own traces a line of constant conductance. Adding series inductance simply adds more positive reactance at higher frequencies and skews the constant conductance circle towards the inductive side of the Smith chart, as shown in Figure 6.28. In terms of R and X, also plotted in Figure 6.28, the reasons for the "loop" in Figure 6.25 become obvious, when it is noticed that the equivalent series resistance first increases and decreases, while the equivalent series reactance increases, decreases, and increases again as we pass through resonance.

This exercise illustrates one of the most important, yet least well-known, tricks in oscillator design. The device in Figure 6.25 is a series representation, because its resistance becomes less negative with increasing drive level A. [We are plotting the negative of Z_D, that is, $Z(A)$, to deal with positive entities.] The rule is this: Series-type devices, in which the resistance becomes less negative with drive, should be terminated with series resonant circuits. The converse is also true, as illustrated above, and series-type devices should not be terminated with parallel-resonant circuits, since, as we have just seen, these can offer multiple frequencies at which the equations for steady-state oscillation (6.14) are satisfied. Whenever this occurs, the potential for mode hopping and hysteresis exist.

It goes without saying that the dual also applies. When the impedance variations of shunt-type devices are plotted either on the Smith chart or on Cartesian coordinates with conductance-susceptance (G-B) axes, the device conductance becomes less negative with drive. Assuming its reactive part varies little, the device line $Y(A)$ will plot from right to left on the G-B plot. In this case, which is the dual circuit of Figure 6.25, it is important to terminate such a device with a parallel resonant circuit, so that the

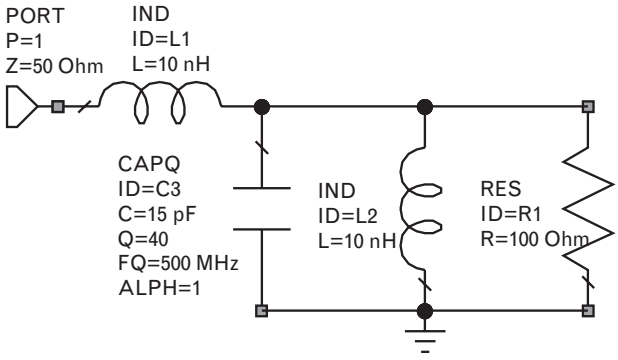
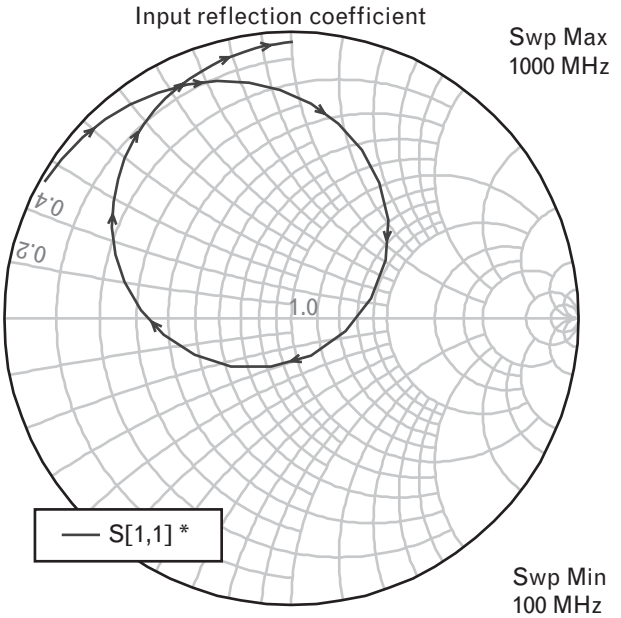


FIGURE 6.27
A simple parallel RLC circuit and a series inductance.



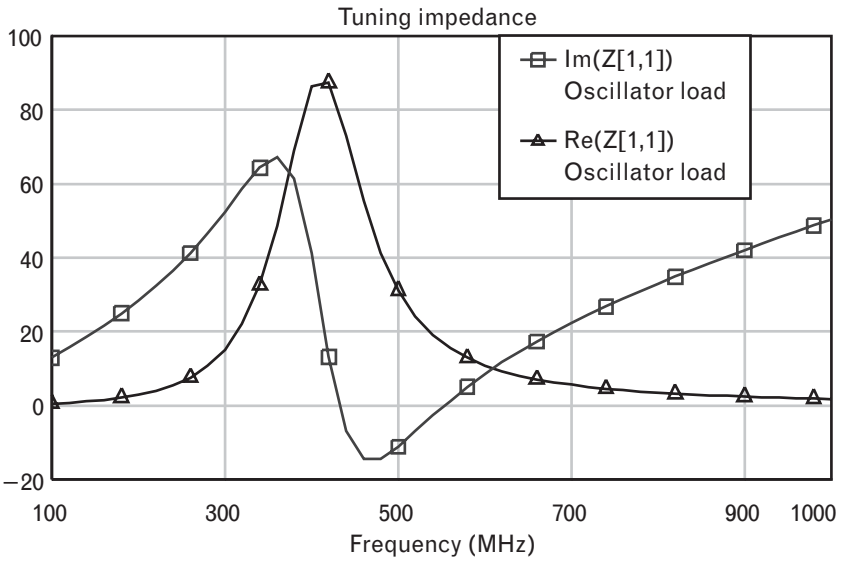


FIGURE 6.28
The equivalent series resistance and reactance of the circuit of Figure 6.27, showing a "loop" characteristic through resonance.

equivalent shunt conductance of the load is constant with frequency and there is only one potential oscillation point as the device is tuned. The crossing angle is then 90° and phase-noise is once again minimized.

We can see now why a 90° crossing angle is important: any fluctuation in load reactance is translated into a frequency shift as the new intersection point moves along the device line. If the crossing angle is ϕ rather than 90°, the same change in reactance will increase the frequency shift by an amount $1/\sin(\phi)$ compared to the shift when the crossing angle is 90°. In other words, the effect that any AM noise on the tuning parameter has on

the oscillation frequency is minimized when the load line and device line are orthogonal, and it is potentially infinite when they are parallel.

In many circuits, the load is tuned through a varactor diode. In this case, the load curve $Z_L(f)$ will change with tuning, typically its reactive part more than its real part. This corresponds to keeping the device line $Z(A)$ fixed and changing the point of intersection by varying the load instead. Any noise on the varactor voltage or thermal noise will change the varactor reactance and, through the mechanism just described, will detune the oscillator. Apart from ensuring the correct crossing angle, such noise effects may be minimized by using multiple smaller varactor diodes (so the noise voltages on each are uncorrelated), and by ensuring the diodes are never driven into forward conduction by the RF voltage across them so that the capacitance itself is stable.

For integrated VCOs, the resonator is rarely incorporated on-chip. Package parasitics can create their own resonant modes, so care is needed to ensure that the oscillator load is of the form desired. Parallel tank circuits, in which a varactor is used in shunt across an inductive load, are more susceptible to spurious resonances than series L-C circuits, for which the package parasitics are more easily absorbed into the intended load. The parasitics can also limit the tuning range of the resulting circuit. The principles outlined above are still the same when $Z_L(f)$ is tuned, and it is important in one-port oscillator design to ensure:

1. That the device behavior is characterized as series or shunt, in which the resistance or conductance, respectively, becomes less negative at larger power levels;
2. That the load impedance presents a single, appropriate termination to the device. For a series device, this will be like a series RLC circuit; for a shunt device, this will be like a parallel RLC circuit;
3. That the intersection point between the device line $Z(A)$ and the load line $Z_L(f)$ occurs at the desired frequency and drive level;
4. That the crossing angle between $Z_L(f)$ and $Z(A)$ is 90° for minimum phase noise.

6.1.2.5 Characterization of a one-port oscillator by its Q factors

Figure 6.29 shows the load impedance $Z_L(\omega)$ of a series RLC circuit plotted on R-X axes. The frequency gradations are marked in units of $\Delta\omega = 2\pi\Delta f$. Around resonance, if the reactance of the load changes by an amount ΔX, we would expect from our previous analysis of tuning that a load with a larger change in ΔX for a given frequency step would yield an oscillator more tightly locked to the oscillation frequency than would a load with a smaller change in reactance.

In fact, for a series resonant circuit we have

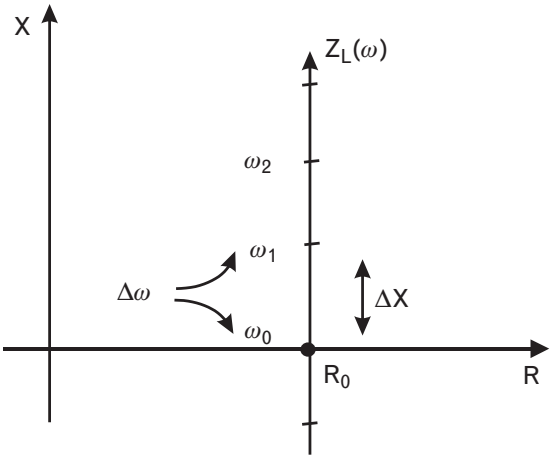


FIGURE 6.29
The load impedance for a series resonant circuit with frequency.

$$X = \omega L - \frac{1}{\omega C} \tag{6.28}$$

so by differentiating we may write at *resonance*

$$\left.\frac{dX}{d\omega}\right|_{\omega_0} = L + \frac{1}{\omega_0^{\,2} C} = 2L \tag{6.29}$$

and using the fact that the Q of a series resonant circuit is $\omega_0 L/R$, we may write

$$\left.\frac{dX}{d\omega}\right|_{\omega_0} = \frac{2QR}{\omega_0} \tag{6.30}$$

or alternatively

$$Q = \frac{\omega_0}{2R}\left.\frac{dX}{d\omega}\right|_{\omega_0} \tag{6.31}$$

Some caution is needed in applying this equation because the frequency of oscillation is not always the resonant frequency. Rather, it is the frequency at which the resonant circuit adds sufficient phase with other feedback elements to make the total phase around the loop equal to zero.

Referring to Figure 6.29, a higher Q circuit will have a larger reactance slope, or a larger reactance change per unit frequency, than a lower Q circuit. Any fluctuation in device reactance (due to noise or temperature) in a high-Q oscillator circuit will result in less frequency shift than in a low-Q oscillator circuit, because from (6.31)

$$\Delta\omega = \omega_0 \frac{\Delta X}{2RQ} \tag{6.32}$$

We can write an alternative expression for Q that eliminates the real part of the impedance by noting that dX/R is just the ratio of the change in reactance around resonance to the real part of the impedance, or tan $\Delta\theta$, where $\Delta\theta$ is the phase angle of the impedance. For small perturbations around resonance, tan $\Delta\theta \approx \Delta\theta$, so we may write (6.31) as

$$Q = \frac{\omega_0}{2} \frac{d\theta(\text{rads})}{d\omega}\bigg|_{\omega_0} = \frac{\pi}{360} f_0 \frac{d\theta(\text{degs})}{df}\bigg|_{\omega_0} \tag{6.33}$$

which is sometimes easier to simulate or measure.

The concept of oscillator Q has proven useful for all types of oscillators, particularly to compare different circuit loads. As before, we break an oscillator into its device part and its load, but we further decompose the load into a resonant circuit with its losses, and an output load resistor, as shown in Figure 6.30.

Associated with (6.31) we may define three different load resistors, and consequently, three types of Q.

The first Q is that associated with the device itself, and it is known as the "loaded Q" or Q_L of the oscillator. It is defined as

$$Q_L = \frac{\omega_0}{2|R_D|} \frac{dX}{d\omega}\bigg|_{\omega_0} = \frac{\omega_0}{2(R_R + R_L)} \frac{dX}{d\omega}\bigg|_{\omega_0} \tag{6.34}$$

because the total resistance seen by the device is that associated with the resonant terminating circuit and the output itself. Q_L is used in calculating the phase noise of the oscillator.

The second Q is that seen by the load and is known as Q_{EXT} since it is measured at the external terminals of the oscillator itself. It is given by

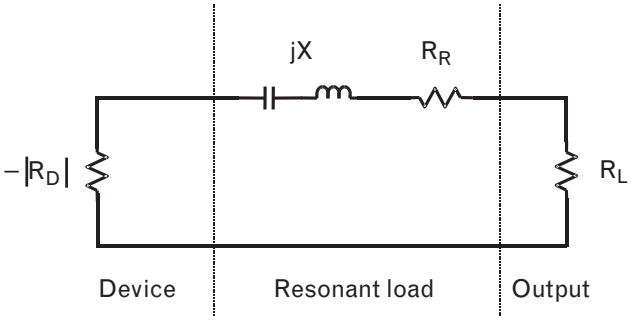


FIGURE 6.30
An oscillator considered as a device, a resonant load, and an output load.

$$Q_{EXT} = \frac{\omega_0}{2\left(-\left|R_D\right| + R_R\right)} \left.\frac{dX}{d\omega}\right|_{\omega_0} = \frac{\omega_0}{2R_L} \left.\frac{dX}{d\omega}\right|_{\omega_0} \tag{6.35}$$

Q_{EXT} is a measure of how tightly locked the frequency of oscillation is with a particular load resistance, and it determines the frequency pulling characteristics of the oscillator through

$$\Delta f_{peak-peak} = \frac{f_0}{2Q_{EXT}} \left(VSWR_L - \frac{1}{VSWR_L} \right) \tag{6.36}$$

Using this expression, we may calculate Q_{EXT} by measuring the peak-to-peak frequency change of the oscillator about the oscillator frequency f_0 when it is pulled with a load of reflection coefficient of $VSWR_L$ through all its angles. For instance, this load can be achieved with an attenuator of x dB in series with a *sliding short*, which is a short-circuited transmission line of variable length. As the line length is varied, its impedance changes in a circle around the outside edge of the Smith chart, through angles of zero to 2π corresponding between zero and a half-wavelength of line length. The effect of the attenuator increases the return loss by twice the value of attenuation. The impedance seen by the oscillator then traces a circle on the Smith chart but at a diameter corresponding to this return loss. Frequency pulling can also occur unintentionally in oscillators when the output power amplifier causes the oscillator to become *injection locked* through feedback. This can especially occur when the oscillator and power amplifier are on the same substrate.

The third Q is Q_O and is just the Q corresponding to the resonator. This is typically known from data sheets or from direct measurements. We may write

$$Q_O = \frac{\omega_0}{2R_R} \left.\frac{dX}{d\omega}\right|_{\omega_0} \tag{6.37}$$

In all three expressions for Q we use the same value for $dX/d\omega$. Therefore, if two of the three expressions are known or can be measured, we can deduce the third using

$$\frac{1}{Q_L} = \frac{1}{Q_O} + \frac{1}{Q_{EXT}} \tag{6.38}$$

This equation shows the trade-off inherent in *coupling* a resonator circuit more or less tightly with a negative resistance device. The easiest way

to consider coupling is to model it as an impedance transformation, as shown in Figure 6.31, where a resonator is coupled into a negative resistance device through a transformer.

The condition for steady-state oscillation in such an oscillator model is that the device resistance is equal and opposite to the load resistance plus the equivalent series resistance of the resonator, seen through the transformer. Increasing the step-up ratio of the transformer not only increases the equivalent series resistance of the resonator seen by the device, but also the reactance slope $dX/d\omega$ seen by the circuit at resonance as well. Consequently, to ensure oscillation, the load resistance will need to be reduced to compensate for the increased resonator resistance. However, the values of Q will also be changed because of the change in reactance slope. In effect, by increasing the turns ratio of the transformer, we can more closely couple the resonator into the oscillator circuit to trade off the relative values of Q_L, Q_{EXT}, and Q_O for improvements in noise, power, or frequency pulling. Of course, Figure 6.31 is a model only. In practice, the coupling of a resonator into a circuit can be altered in a number of ways. Historically, the notion of coupling came about at microwave frequencies, where the dimensions of an iris could be changed to change the coupling of a resonant waveguide with a Gunn diode oscillator. On microstrip, the coupling can be increased by moving a dielectric resonator closer to the microstrip line. At RF frequencies, the coupling can be changed by altering the size of a coupling capacitor in series with the crystal or other resonator.

6.1.3 Transistor oscillator configurations

To this point, the one-port design approach has assumed that we have synthesized a negative resistance or negative conductance, either through ensuring that the Barkhausen criterion is met when feedback is applied, or through known configurations of circuits in which the transistor is unstable. In all cases, oscillator circuits will have some feedback from output to input, even though it may not be immediately identifiable.

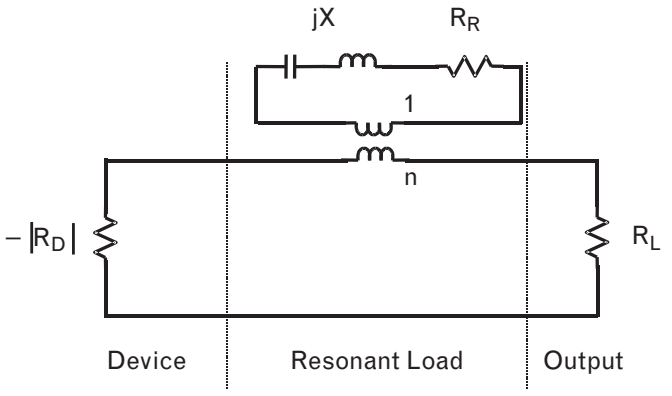


FIGURE 6.31
A resonant load coupled in series with a negative resistance device through a transformer, to create an oscillator.

6.1.3.1 Generalized L-C oscillator topologies

Common-emitter and common-source transistors, with their reverse-biased junction between base and collector, or gate and drain, will have only small capacitive feedback between the output and input. At low frequencies, the phase through any added external feedback network would have to be 180° to achieve the total loop phase shift of 360° necessary for oscillation. The transistor capacitive feedback on its own is generally insufficient to sustain oscillation, as indicated by the stability figures of most of these devices.

Instead, devices in which the much larger base-emitter or gate-source capacitance is used as the feedback element between the output and input are generally more unstable. These configurations occur in common-collector or common-drain connections. Such connections have found general use in wideband VCOs because of their ability to generate a negative resistance across a broad range of frequencies. This requires proper termination of the input base or gate, and matching of the output emitter or source to null out the reactance presented by the device and to select the desired operating frequency. The VCO example at the end of this chapter illustrates how an arbitrary oscillator topology can be created using a common-collector device. In principle, such an approach is common in the microwave frequency ranges, where the assumptions underlying more direct and specific topologies, such as the Colpitts, no longer apply.

A second mechanism to generate negative resistance is to use the transistor in common-base or common-gate configuration, normally adding series inductance in the common (base/gate) terminal in order to maximize the negative resistance or negative conductance seen at the collector or drain. Generally, the emitter or source will have a capacitive termination in order to achieve this. The phase shift through the feedback network must be 0° in this configuration since there is minimal phase shift through the transistor itself. Again, the input and output terminations are usually specifically configured for the particular device to achieve this.

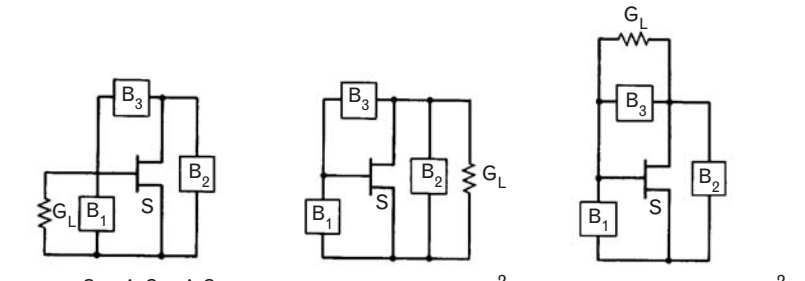
The art of transistor oscillator design using such configurations involves tuning the reactance at the input port of the unstable device to present maximum negative resistance or conductance looking into the remaining output terminal. On the Smith chart, it is frequently easier to tune the input so the output reflection coefficient exceeds unity by as much as possible. The design is completed by choosing the appropriate output termination to resonate the circuit and achieve a 90° crossing angle with the device impedance line seen at the output port. We will illustrate this in the VCO design at the end of this chapter.

In integrated circuits, multiple devices can be connected together to create the necessary feedback conditions. Most frequently, a differential

pair of bipolar transistors with tied emitters and their bases cross-coupled capacitively to the opposite collector provides a flip-flop arrangement that is guaranteed to oscillate. The tank circuit can consist of spiral inductors between the collectors and the rail, parallel resonant with MOS varactor capacitors that allow for tuning. FETs can also be used and are more convenient in CMOS. The gate of the FET will self-limit when driven between pinch-off (threshold) and turn-on, and this can be several volts in magnitude. As shown in the following section on phase noise, maintaining a large signal voltage helps to minimize the oscillator phase noise, and in this respect, FET oscillators in CMOS have been reported [8] with comparable phase noise to bipolars, in spite of their higher $1/f$ noise. Circuits of this type are extensively used for applications between 1 and 5 GHz.

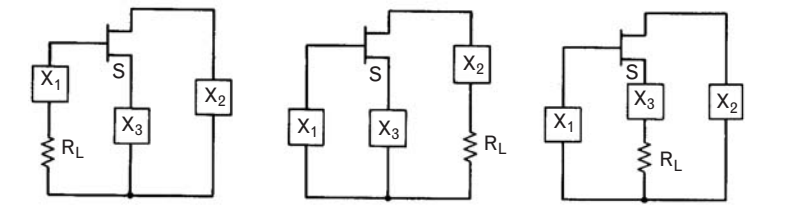
Vendelin et al. [9] summarize a number of other general oscillator configurations in which each of the three nodes of the transistor (FET or bipolar) are terminated in a reactance that is calculated from the *Y*-parameters or the *S*-parameters [10] of the device. These are summarized in Figure 6.32. Large-signal parameters can be used in order to reflect the conditions likely to exist when the device is limiting. The load resistance can be connected at any one of the three nodes, and because the reactances can be connected either in series with a node, or between adjacent pairs of nodes, there are a total of six configurations that can be designed. This is an analytic technique that enables the embedding circuit to be calculated automatically from a series of formulae. Furthermore, it selects the load resistor and device gain to automatically maximize the power into the load. However, such techniques seem to be rarely used for a number of reasons. First, the load resistor is a design variable itself, and its value must be specified by the design equations rather than freely chosen. Second, the resulting circuit requires reactive terminations at, or between, all three nodes, and this is not always necessary in other types of configuration. Finally, although the circuit is designed to oscillate at the specified frequency, the implementation of the specified reactances as real inductors and capacitors means that extra consideration and analysis is still required to ensure that the crossing angle is correctly chosen. The resulting oscillator embedding circuit, specified in terms of reactance at a single frequency, imposes no constraint on the frequency locus of the load or the Q of the circuit. In reality, it will be necessary to designate one of the terminating reactances as a resonator and shape its impedance characteristics accordingly to ensure proper frequency locking and phase noise.

At frequencies below about 1 GHz, more direct approaches can be used to design an oscillator, such as the Colpitts topology and its variants. By using the Colpitts template for oscillator design, the designer is assured of a circuit that avoids the problems of mode hopping and hysteresis.



G_L	$C_1 + A_rC_3 + A_iC_4$	$(C_1+A_rC_3+A_iC_4)/\|A\|^2$	$(C_1+A_rC_3+A_iC_4)/\|A-1\|^2$
B_1	$C_2 + (1-A_r)(C_4+C_3A_r/A_i)$	$C_1(A_r-1)/A_i+C_2$	$(C_1+C_3)A_r/A_i+C_2+C_4$
B_2	$C_3(A_r-1)/A_i+C_4$	$[C_1(A_r-\|A\|2)/A_i-A_iC_3+A_rC_4]/\|A\|^2$	$-(C_1+C_3/A_i)$
B_3	$-C_3A_r/A_i-C_4$	C_1/A_i	$[C_1-(1-A_r)G_L]/A_i$

where: $C_1=-Re(Y_{11}+AY_{12})$, $C_2=-IM(Y_{11}+AY_{12})$, $C_3=-Re(Y_{21}+AY_{22})$, $C_4=-IM(Y_{21}+AY_{22})$, $A=A_r+jA_i$, $-(Y_{21}+Y_{12*})/2\,ReY_{22}$



R_L	$D_1+F_rD_3+F_iD_4$	$(D_1+F_rD_3+F_iD_4)/\|F\|^2$	$(D_1+F_rD_3+F_iD_4)/\|1+F\|^2$
X_1	$D_2-(1+F_r)(D_4+D_3F_r/F_i)$	$D_1(1+F_r)/F_i+D_2$	$(D_1-D_3)F_r/F_i+D_2-D_4$
X_2	$-D_3(1+F_r)/F_i-D_4$	$(D_1(F_r+\|F\|^2)/F_i-F_iD_3+F_rD_4)/\|F\|^2$	$(D_1-D_3)/F_i$
X_3	D_3Fr/F_i+D_4	$-D1/Fi$	$[(1+F_r)(D_4-(D_1-D_3)F_r/F_i)$ $-F_iD_1]/\|1+F\|^2$

where: $D_1=-Re(Z_{11}+FZ_{12})$, $D_2=-Im(Z_{11}+FZ_{12})$, $D_3=-Re(Z_{21}+FZ_{22})$, $D_4=-Im(Z_{21}+FZ_{22})$, $F=F_r+jF_i$, $F=(Z_{21}-AZ_{11})/(AZ_{12}-Z_{22})$

FIGURE 6.32 *Embedding circuits for oscillation, designed to give optimum power into the load resistor shown. (From: [9]. © 1990 John Wiley & Sons, Inc. Reprinted by permission.)*

6.1.3.2 The Colpitts oscillator and its variants

The analysis of a Colpitts oscillator is very simple. The basic ac configuration is shown in Figure 6.33, together with a nearly trivial low-frequency model for the circuit. The circuit itself consists of a tapped capacitor stack between the base and collector, with feedback to the emitter via the

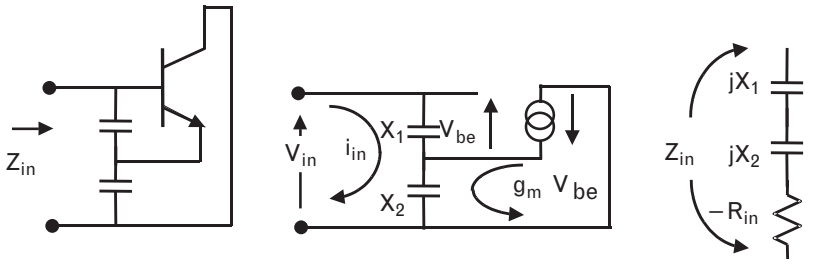


FIGURE 6.33 *The basic Colpitts oscillator configuration, and the equivalent low-frequency model.*

capacitive tap. The resonant circuit is connected across the two capacitors and will be inductive at the oscillation frequency. If we assume that the reactance of X_1 is so low that it totally dominates the base resistance and base-emitter capacitance, and model the collector as a current source of magnitude $g_m v_{BE}$, then the voltage across the capacitor stack is given by

$$\begin{aligned} v_{IN} &= i_{IN}\left(jX_1 + jX_2\right) + g_m v_{BE}\left(jX_2\right) \\ &= i_{IN}\left(jX_1 + jX_2\right) + g_m\left[i_{IN}\, jX_1\right]\left(jX_2\right) \\ &= i_{IN}\left[-g_m X_1 X_2 + j\left(X_1 + X_2\right)\right] \end{aligned} \tag{6.39}$$

The input impedance looking between the base and the collector is thus

$$Z_{IN} = \frac{v_{IN}}{i_{IN}} = -g_m X_1 X_2 + j\left(X_1 + X_2\right) \tag{6.40}$$

This fundamental result for the Colpitts configuration shows that the Colpitts input impedance consists of a negative resistance $-g_m X_1 X_2$ in series with a reactance equal to the series reactance of the two capacitances. The negative resistance arises in (6.39) because the product of jX_1 and jX_2 in the final term gives rise to a $j^2 = -1$ term. In fact, the circuit would work equally well if both X_1 and X_2 were inductive reactances, because the j^2 term will still arise as long as both X_1 and X_2 are of the same sign. Later, we will deploy the fact that if one is inductive and the other capacitive, then the resistance is positive.

The Colpitts topology can also be analyzed by calculating its open-loop voltage gain. Using Figure 6.33, it is straightforward to apply an input voltage at the base and to calculate the resulting open-loop voltage across the output X_1 when a load impedance Z_L is connected between the base and the collector. We can show that

$$AL(j\omega)H(j\omega) = -g_m \frac{jX_1(\omega)\, jX_2(\omega)}{jX_1(\omega) + jX_2(\omega) + Z_L(\omega)} \tag{6.41}$$

The conclusions are the same as above, although an expression for gain such as this is always useful for analyses such as the Nyquist stability plot.

From (6.40) we can see that as the device compresses and g_m becomes smaller, the negative resistance becomes proportionally smaller as well. This, therefore, is a series-type device, so the load impedance, between the base and the collector, needs to be a series resonant circuit. The load resistor needs to be smaller than $g_m X_1 X_2$ so that oscillations will start up from small-signal ($R_L < |R_D|$), and the load reactance needs to equal $-j(X_1 + X_2)$, which is inductive at the oscillation frequency. In the case where both

capacitances are equal, this gives rise to the expression for startup of oscillation:

$$\frac{1}{\omega C} > \sqrt{\frac{R_L}{g_m}}$$

Because the derivation for the negative resistance assumes that the output is a current source that generates a current $g_m v_{BE}$, we can in theory insert any impedance into the collector loop without modifying the equations above. This will be useful when we look at extracting power from the oscillator, because the collector can drive a load resistance without modification of the basic negative resistance equations just derived.

It is also sometimes useful to insert a small series resistor with the emitter. Because of the negative feedback it introduces, this can reduce the phase noise of the oscillator as it desensitizes the negative resistance to changes in g_m. The impact of a series resistor R_E is to increase the effective emitter resistor of the device to $r'_E = r_E + R_E$, where r_E is the emitter resistor of the unloaded transistor and r'_E is the new "emitter" resistance of the device extended to include the added resistor. Thus, the new g_m of the device g'_m is given by

$$g'_m = \frac{1}{r'_E} = \frac{1}{r_E + R_E} = \frac{1}{\frac{1}{g_m} + R_E} \tag{6.42}$$

and the new Colpitts input impedance by

$$Z'_{IN} = -\frac{X_1 X_2}{\frac{1}{g_m} + R_E} + j\left(X_1 + X_2\right) \tag{6.43}$$

Although this has the effect of reducing the negative resistance presented at the base-collector port, this resistance can now be made more stable to fluctuations in device g_m through temperature, drive, or device variation effects, if R_E is chosen to be much greater than $1/g_m$.

6.1.3.3 The Clapp, Pierce, and Hartley variants of the Colpitts oscillator

This basic circuit provides the ability to create a number of different circuit variations and oscillator types, summarized in Figure 6.34. By grounding the emitter instead of the collector, a Pierce oscillator can be created from the basic Colpitts configuration. A grounded-emitter device is typically easier to bias than one that is common-collector. Furthermore, the bias

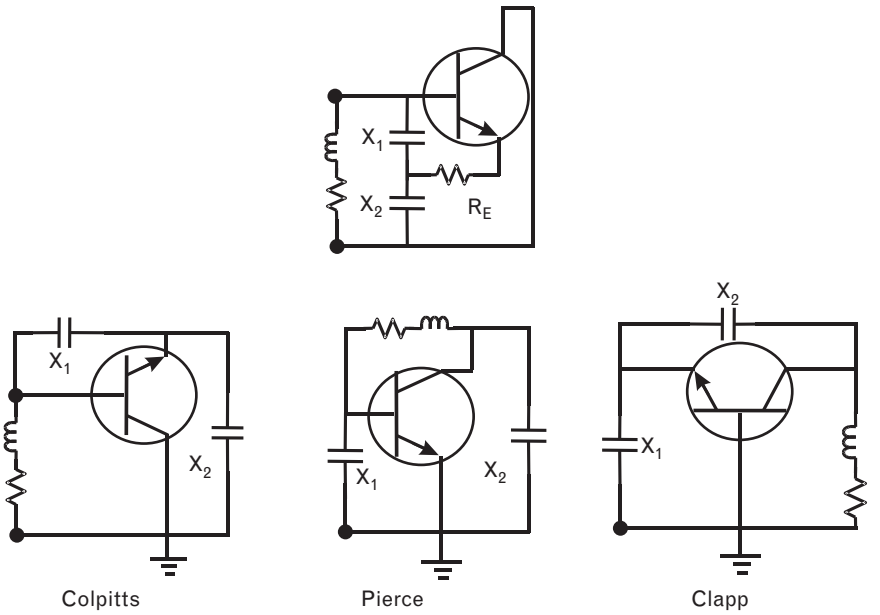


FIGURE 6.34
The Colpitts oscillator and its basic variants.

resistor network loads only X_1 in the Pierce oscillator, rather than $X_1 + X_2$ as in the Colpitts, resulting in a higher Q oscillator.

By choosing X_1 and X_2 to be both inductive reactances, rather than capacitive, the Hartley configuration results.

The Clapp oscillator has rather inconsistent definitions in the literature. Some authors refer to the Clapp oscillator as a Colpitts configuration with grounded base. Others consider the Clapp, or Clapp-Gouriet oscillator, to be the basic Colpitts configuration externally loaded by an inductor in series with a capacitor. The series capacitance shifts the resonant frequency to a (higher) frequency at which the net resonant frequency of the series inductor and capacitor equals $-j(X_1 + X_2)$. From (6.29) above, the reactance slope of the series LC load at resonance is double the reactance slope of a single inductor, and thus the Clapp oscillator has a higher loaded Q than the Colpitts equivalent. Thus the Clapp is also more stable with frequency; conversely, if the inductor drifts with temperature, the Clapp oscillator will drift at a faster rate than the Colpitts. The extra capacitor can also be implemented as a varactor diode and used for tuning.

At low frequencies, the inductive resonator is typically chosen to be a crystal operating in its inductive region; at RF frequencies, a high Q termination can be achieved by using a ceramic quarter wave transmission line operated in its inductive region; and, of course, for low Q implementations, either a lumped-element or distributed inductor can also be used. The Clapp and Colpitts configurations are simpler to implement since one side of the crystal or transmission line is grounded; on the other hand, any parasitic capacitance across the crystal to ground directly shunts the desired inductance and causes detuning of the oscillation frequency. In this latter

regard, the Pierce oscillator is a better choice because any stray parasitic capacitance will appear across X_1 and X_2 to ground, rather than the crystal, and the parasitics are then swamped by those much larger circuit elements.

For very high frequencies into the RF range and beyond, the common-base configuration of the oscillator is often the preferred choice. The common-base configuration is able to operate at higher frequencies than the common-emitter because the transistor has no voltage gain. In common-emitter, the equivalent loading at the input by the capacitance C_μ between the base and collector is magnified by the large voltage gain appearing across it. This capacitance referred to the input is known as the Miller capacitance, and it lowers the frequency of the dominant input pole that causes the gain to roll off. With a common-base transistor, the absence of any voltage gain means that the input capacitance is essentially unaffected by C_μ so the transistor operation can extend into higher frequencies.

6.1.3.4 Crystal oscillators

As just noted, the resonating inductor can be implemented by choosing a crystal operating in its inductive region, where it is loaded by the Colpitts capacitance X_1 and X_2. Since the effect of any stray capacitance across the crystal and its own package capacitance is to add further load to the crystal resonator, the oscillation frequency will lie above the crystal series resonance, where its total reactance is inductive. These loading capacitances also transform the series loss of the crystal resonator, increasing the loading on the device seen at the Colpitts terminals. These effects are illustrated in Figures 6.35 and 6.36.

Figure 6.35 shows the equivalent circuit of the crystal resonator introduced in Volume I, Chapters 7 and 8. There are two series resonant arms shown, known as *motional* arms: one representing the fundamental resonance, and the other the third overtone. We will use this crystal later to design a 45.455-MHz Colpitts oscillator, to work with a 45-MHz mixer. The motional arm modeling the third overtone frequency of the crystal has a series resonance at 45.455 MHz, and at this frequency the fundamental arm (resonant at 15.1517 MHz) appears as a huge shunt inductor and can be ignored. In Figure 6.35, we see that the point around the third overtone at which the net reactance equals zero has shifted just 38 Hz above the resonance of the motional arm, showing that the loading effect of the 4-pF parasitic capacitance is negligible for a high Q crystal such as this. The effective series resistance of the crystal, 40Ω, also remains approximately equal to the resistance of the motional arm because of the high Q of the crystal. However, Figure 6.36 shows the impact of the series loading of the Colpitts capacitance on the crystal. If we assume a Colpitts load capacitance of 35 pF, we first note that the frequency at which the overall reactance equals zero has now shifted higher by 1,790 Hz, because the crystal must now look more inductive to compensate for the Colpitts loading. This shift

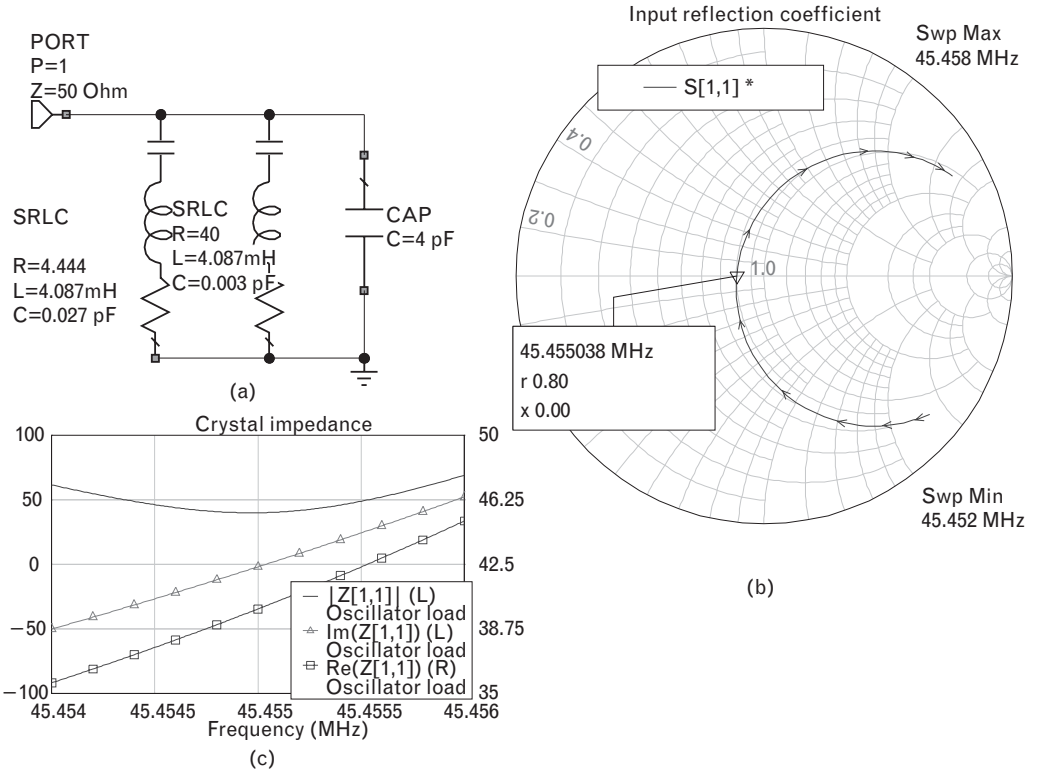
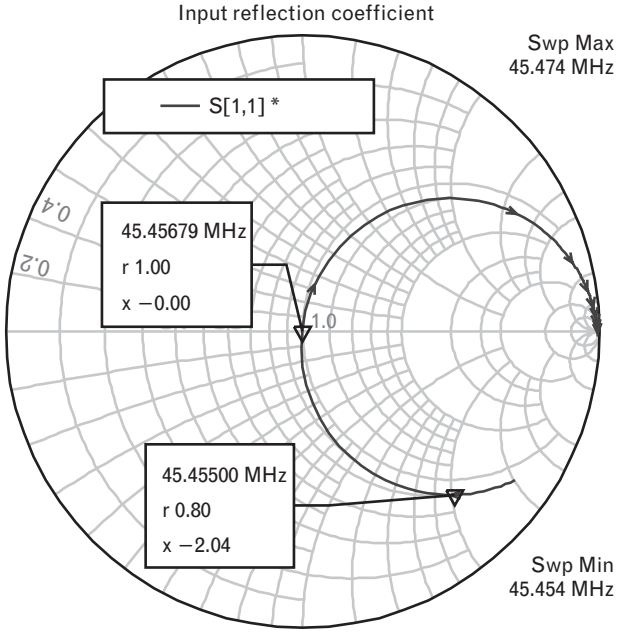


FIGURE 6.35 *(a) The equivalent circuit for a 15.152-MHz crystal; (b) its input reflection coefficient at the third overtone 45.455 MHz; and (c) the real and imaginary parts of its impedance, and magnitude of the impedance.*

FIGURE 6.36
The input reflection coefficient of the crystal when loaded in series with a Colpitts reactance that corresponds to a capacitive loading of 35 pF.



can also be calculated from (6.29), which specifies a reactance slope of $2L$ or 0.0082Ω/rad/second around resonance. Thus, for an incremental series reactance introduced by the 35-pF capacitor, equivalent to $-j100\Omega$ at 45 MHz, the radian frequency shift is $100/0.0082 = 12{,}230$ radians/second or 1,950 Hz, close to that shown on the Smith chart in the figure. Second, the equivalent series resistance of the crystal has increased from 40Ω in Figure 6.35 to 50Ω in Figure 6.36. This is because the package and parasitic capacitances of the crystal itself introduce a complex impedance transformation. The impact of this is that the Colpitts device resistance must be chosen somewhat greater (i.e., more negative) than the crystal motional resistance, not only to allow oscillation to build up but also to overcome the somewhat increased effect of the total crystal load resistor.

To select the third overtone of the crystal, we exploit the fact noted earlier, that if X_1 and X_2 are of opposite sign, their product in (6.39) results in a positive resistance. A selector circuit of the form shown in Figure 6.37 can be used for one of X_1 or X_2.

At low frequencies, this circuit appears inductive, while above resonance, its capacitance dominates. By setting the resonant frequency to be midway between the highest unwanted overtone and the next desired overtone, X_1 and X_2 will be of opposite sign at all lower-frequency

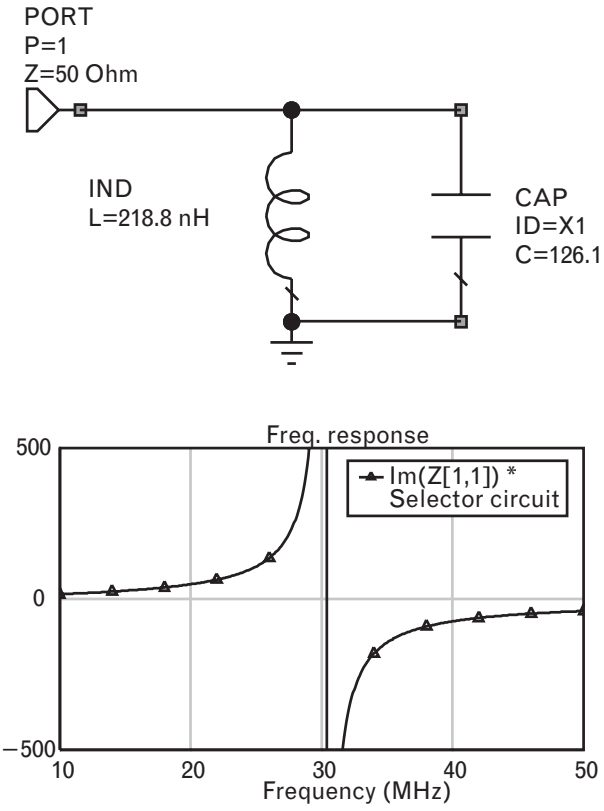


FIGURE 6.37
An L-C selector circuit used to replace either X_1 or X_2, in order to select the desired crystal overtone. The selector reactance with frequency is shown.

unwanted overtones, but of the same sign (capacitive reactance) at the desired overtone above the selector resonant frequency.

This then gives some basic configurations for the Colpitts crystal oscillator, whose RF schematics are shown in Figure 6.38.

Figure 6.38(a) shows the circuit we have so far configured, in which the emitter is connected to the tap of a capacitive stack, and the input between base and collector is loaded with a crystal in the inductive region. An emitter resistor is added to desensitize the impedance to changes in the device transconductance, which also improves phase noise, and the L-C circuit described above is used to select the desired overtone of the crystal. Power is taken from one side of the crystal through a coupling capacitor into an output load resistor. The capacitive loading can be increased by increasing the series capacitor. This reduces the loaded Q of the oscillator and will deteriorate its phase noise and frequency pulling, although the output power will be greater.

Figure 6.38(b) shows an alternative way to couple output power from the oscillator, through the collector. As noted earlier, the collector current is insensitive to its load impedance if it is modeled as a current source, so the collector is semi-isolated and a load impedance can be directly connected at this point. However, any voltage gain from the transistor then loads its base with a large Miller capacitance that can detune the oscillator. Instead, the collector can be terminated with the (small) emitter resistor presented by a second transistor, which keeps both its voltage gain and Miller capacitance low. This (second) device is just a common base amplifier, which decouples the load from the Colpitts device and presents a

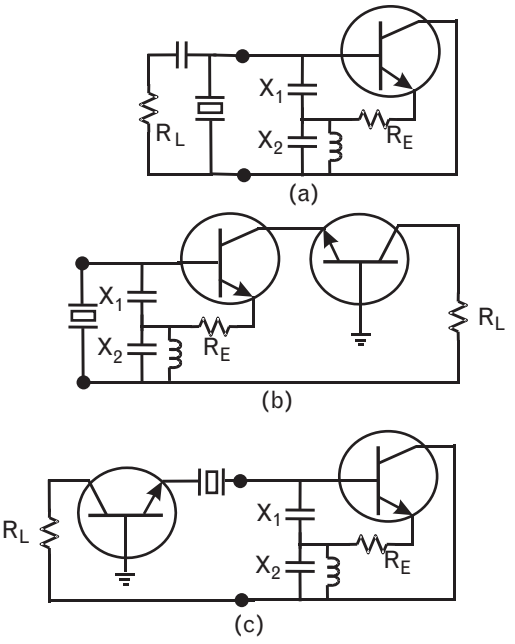


FIGURE 6.38
Basic configurations of the Colpitts oscillator, with a crystal operated inductively as the resonant load: (a) crystal connected to capacative stack and loaded with resistor; (b) an alternative way to couple output power from the oscillator, through the collector; and (c) the output load resistor is effectively in series with the crystal.

constant output impedance to the collector. The two transistors form a cascode connection, which has low overall Miller capacitance and good isolation at the output. The crystal is thus not loaded with parasitic or output elements in this configuration, so the loaded Q is higher.

A variation of this configuration is when the output load is tuned to a harmonic of the fundamental oscillation frequency. This can be done because there is good isolation between the harmonic output frequency and the fundamental at the input. The load impedance at the collector of the Colpitts device can then be kept low at the fundamental frequency to reduce the loading effect on the oscillator itself.

Figure 6.38(c) shows a configuration in which the output load resistor is effectively in series with the crystal. From the oscillator side, the Colpitts impedance is terminated by the crystal in series with the emitter resistor r_E of a second device; that device, of course, is just a common base amplifier. The load is then connected to the output of that buffer amplifier. In this case, the current in the load is the same as the current in the crystal, so that any noise is very bandlimited by the crystal.

Figure 6.39 shows other RF configurations of the Colpitts using a resonant tank circuit to terminate the device. This would yield a low-Q circuit, since the Q of the tank circuit cannot match the Q of a crystal; however, a crystal is now used in series with the emitter to close the loop to the capacitive tap. At resonance, the crystal appears as a series resistance and the loop is effectively closed with the crystal acting as a series emitter resistor that shifts the magnitude and phase of the g_m of the transistor. These circuits are useful for high overtone operation, because the frequency of

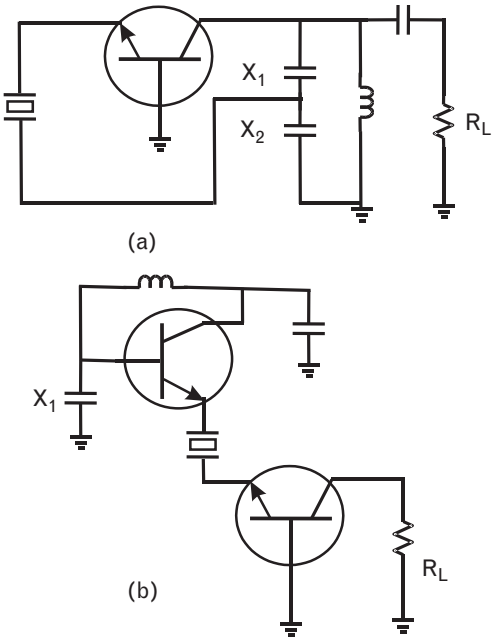


FIGURE 6.39
Colpitts oscillators terminated by an R-L tank circuit, and in which the crystal at resonance is used to close the feedback loop.

oscillation is set by the frequency at which the tank circuit inductance cancels the Colpitts capacitance, and this is set close to the desired overtone frequency. At lower frequencies, although the loop is closed by one of the other series resonant arms within the crystal, the tank circuit is not sufficiently inductive to meet the conditions for oscillation in (6.14).

In the first configuration shown, the output power is capacitively coupled to a load resistor across the tank circuit [this circuit is similar to Figure 6.38(a)]. This is known as the Butler oscillator, and with the transistor connected in common-base, it is capable of high-frequency, low-noise operation. In the second configuration, the load resistor is connected to the collector of a common base amplifier whose current is the same as the current in the crystal. With the capacitive tap at ground (Pierce configuration), the loop is closed when the crystal resistance at resonance in series with the emitter resistor of a second device completes the ground loop to the emitter of the Colpitts device. This circuit is again low noise because the output current is bandlimited by the crystal itself.

6.1.3.5 Examples of Colpitts-class oscillators

A question we frequently encounter with these configurations is: How are they biased? We will illustrate this by presenting several commercial Colpitts-class oscillators as examples. At the same time, we will show some typical circuits used to tune their frequency.

Figure 6.40 shows a 500-MHz Colpitts-type circuit in which a ceramic quarter-wave transmission line is used to load the Colpitts negative resistance and lock the frequency within the desired range. These silver-plated quarter-wave TEM-mode resonators are available in a range of frequencies from 400 to 4,500 MHz, and because the ceramic is of a high dielectric constant material, they are not excessively large. Dielectric constants ranging from 20 to 90 can be used, with different temperature coefficients, to compensate for frequency drift of the device itself with

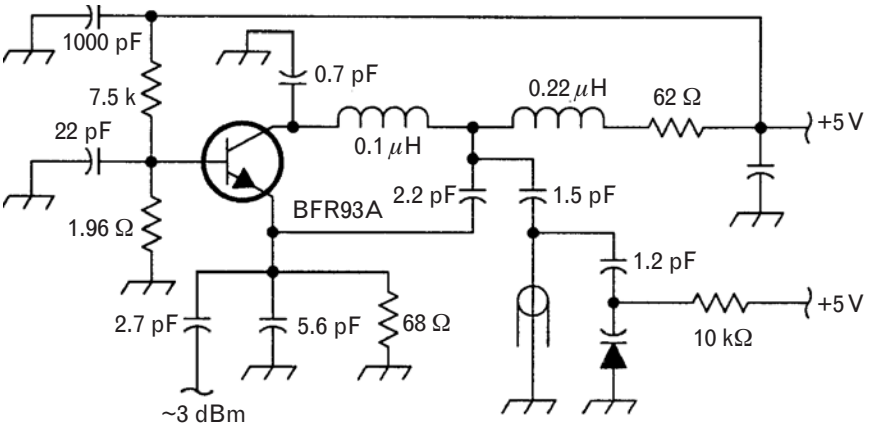


FIGURE 6.40
A 500-MHz oscillator circuit using a ceramic resonator. (From: [11]. © 1994 QEX. Used with permission.)

temperature. The resonator unloaded Q will be several hundred, usually less than 1,000. The example shown is a Clapp configuration, because the base is at RF ground by virtue of the large 22-pF capacitor. Thus, X_1, between the base and emitter, is the 5.6-pF capacitor; and X_2, between the emitter and collector, is the 2.2-pF capacitor. This circuit has a collector impedance in series with the collector, but as already noted, this does not affect circuit operation since the only requirement for (6.39) is that the collector current flow through X_2. In this instance, the 0.1-μH inductor and 0.7-pF capacitor present a 600-MHz lowpass filter at the collector terminals, so it is used to prevent parasitic oscillations at higher frequencies propagating around the feedback loop. The resonator is connected between the base (ground) and the collector. The resonator circuit consists of a series capacitor (1.5 pF) used to adjust the coupling of the resonator, a dc block (1.2 pF), and the quarter-wave line shunted by a varactor capacitor for tuning. We shall examine the structure of this resonant circuit in more detail shortly. Power out from the oscillator is through a coupling capacitor at the emitter. The remainder of the circuit elements are for bias, which is reasonably straightforward.

Figure 6.41 shows a Hartley oscillator using a JFET, in which X_1 and X_2 are inductors. The circuit can be tuned between 74 and 105 MHz using varactor diodes. The Hartley configuration is readily detectable in the figure because the feedback through the autotransformer between the source and the gate of the FET is easily identified. A dc blocking capacitor of 470 pF is required. The drain is grounded through a 22-μF capacitor in parallel with a

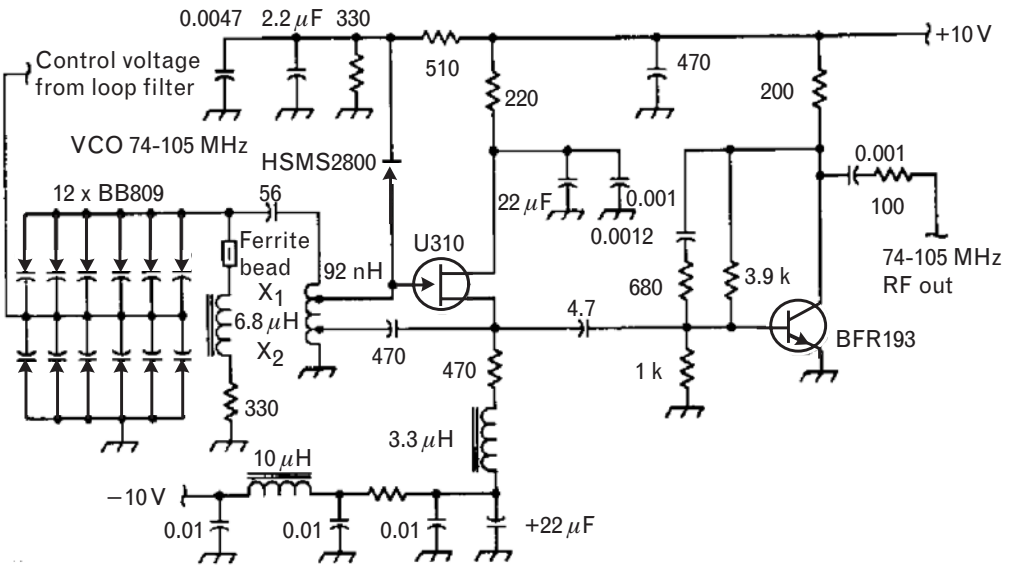


FIGURE 6.41 *A Hartley oscillator, with wideband tuning between 74 to 105 MHz. (From: [12]. © 1994 QEX. Used with permission.)*

1,000-pF capacitor. The second capacitor is used because the self-resonant frequency of the first capacitor is quite low, and at higher frequencies has a relatively large series inductance that will not be an effective ground. The gate is at dc ground and a split supply is used to bias the drain and source, with the source requiring an RF choke for bias because it carries the output signal. The resonant circuit required for a Hartley oscillator is capacitive, and this is achieved using a number of small varactor diodes connected between the gate and drain (ground). This gives improved phase-noise performance [11, 12] because each diode has a smaller capacitance and less noise voltage than a single larger diode, and the individual noise voltages across each diode are uncorrelated. In addition, connecting the diodes back-to-back as shown in the figure will eliminate even harmonics [13], and provide a four-fold increase in RF voltage handling capability. Any forward conduction of the diodes on current peaks would otherwise cause an impulsive change in the varactor bias voltage, giving rise to various harmonics. The designers have also connected a diode between the gate and positive supply rail to prevent excessive peaks of the signal at the gate degrading the phase noise. The second transistor is a small-signal bipolar transistor that serves as a buffer amplifier for the signal. This component is ac coupled to the oscillator, and uses R-C feedback between the drain and base to improve the input match and flatten the gain.

Figure 6.42 shows a VCO again using a quarter-wave ceramic transmission line between the collector and base (ground) as the resonator. The transistor is a Philips Semiconductors BFQ67, a versatile BJT with f_T of 8 GHz. In this example, we used a bias voltage of 4V and 2-mA collector current. Because the base is RF grounded through a 1-μF and 10-pF

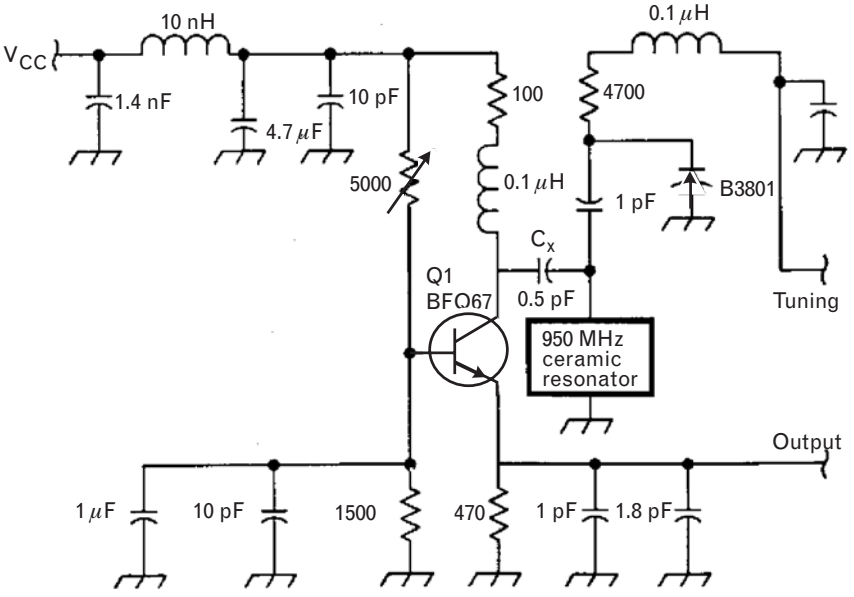


FIGURE 6.42
800-MHz VCO using a Clapp configuration. (From: [12]. © 1994 QEX. Used with permission.)

capacitor combination, this is a Clapp oscillator. X_1 is formed from the combination of the 1-pF and 1.8-pF capacitors, and X_2 is the internal output capacitance of the device itself, between the collector and emitter. No external X_2 is required. The base is biased through a resistive divider; a large emitter resistor of 470Ω improves the thermal stability of the device and the oscillator phase noise; and the collector is fed through an RF choke. The output signal is taken across the emitter of the device.

The resonant circuit is of the same topology as the circuit in Figure 6.40. C_X is the resonator coupling capacitor, used to adjust the loaded Q. It achieves this in the same way as the coupling transformer of Figure 6.31, effectively stepping up both the resistance, the reactance, and the reactance slope of the resonator at the input as the capacitance is reduced. It also serves as a dc block. The resonator itself is a short-circuited quarter-wave line at 950 MHz, so at the oscillation frequency near 800 MHz, below the line resonance, it appears inductive. Tuning is achieved with a varactor that adds shunt capacitance. A 1-pF dc blocking capacitor is necessary between the short-circuited transmission line and the varactor itself, which is tuned through an RF choke and series resistor.

A linear analysis of the impedance seen looking into the device is shown in Figure 6.43. The Smith chart confirms that its impedance lies outside the unit circle with a reflection coefficient greater than one (i.e., the input impedance contains a negative real part). The real and imaginary values of this impedance are plotted in Figure 6.44, together with the total magnitude of the impedance.

As anticipated for the Colpitts configuration, a very broadband negative resistance is achieved between at least 100 MHz and 2 GHz. This confirms the usefulness of this topology. At 800 MHz, we see that for steady-state oscillation the resonator loading the device will need to provide a resistive load smaller than 49Ω in series with an inductance of $+j341\Omega$.

Figure 6.45 shows the nature of the resonant load that is connected to the collector of the Clapp oscillator. The circuit topology looks suspiciously like a shunt R-L network, when, in fact, the Colpitts topology requires a series R-L circuit as its load to achieve a 90° crossing angle and to avoid multiple oscillatory modes. However, the load impedance plot shows that across this range of frequencies the impedance lies on a line of almost constant resistance, so that the oscillator is indeed correctly terminated with a series R-L circuit. Depending on the value of the varactor capacitance used, the resonant circuit can be made inductive at frequencies above about 750 MHz, in series with a resistance that is less than the required 49Ω. The load does, in fact, also pass through a shunt resonance around 950 MHz, where the quarter-wave line appears as an open circuit. However, because the circuit is used below this frequency in its inductive region, in much the same way as a crystal is used at lower frequencies below its parallel resonance, there is the possibility of only a single

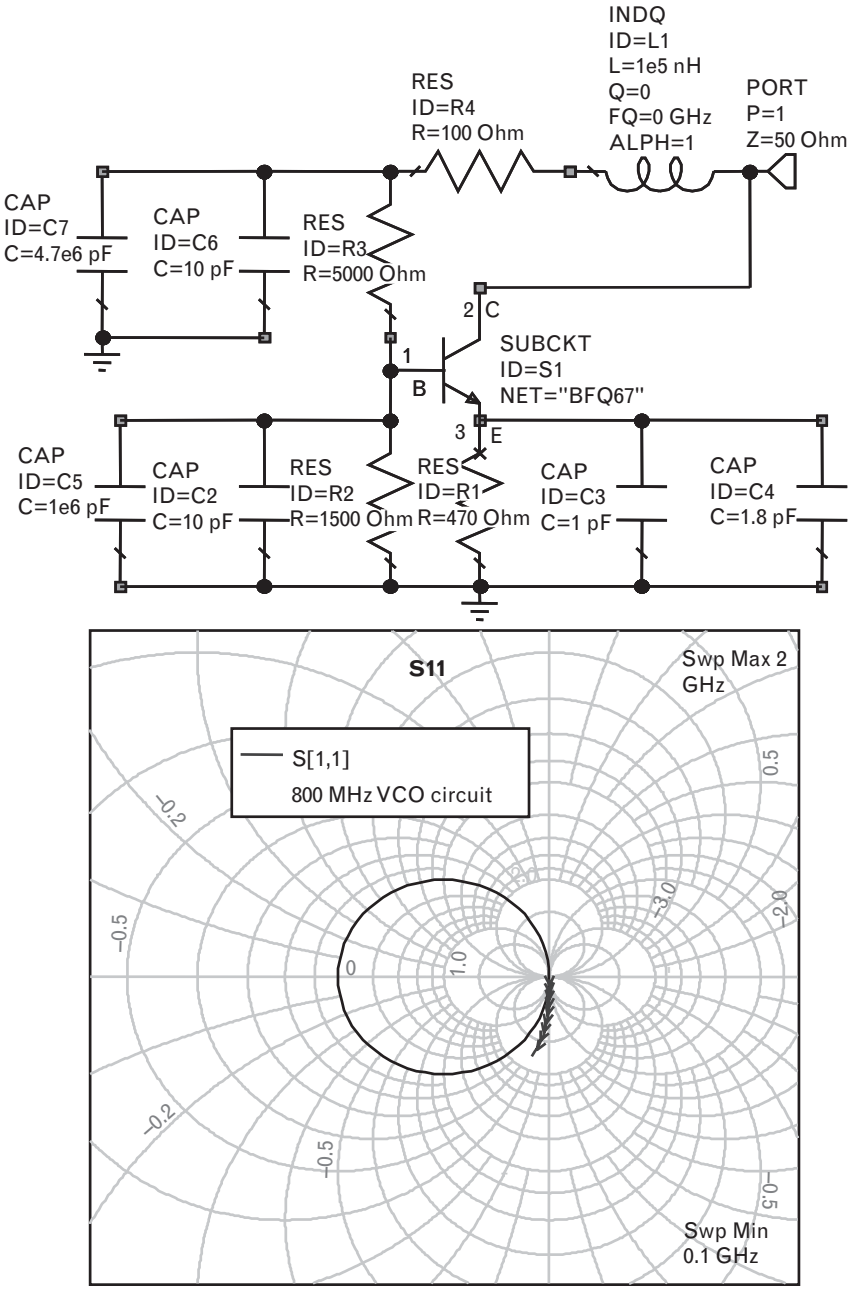


FIGURE 6.43
(a) The device model for the oscillator of Figure 6.42; and (b) the impedance looking into the device output.

oscillation frequency. The load has an inductance of $+j341\Omega$ at just one frequency, depending on the value of the varactor tuning.

This is illustrated in more detail in Figure 6.46, which shows the reactance variation of the resonator around 800 MHz when the varactor capacitance is set to 2.82 pF. As the varactor capacitor is tuned between 1 pF and 3 pF, the frequency at which the inductance of the resonator cancels out the net capacitance of the Clapp device ($-j341\Omega$) changes from

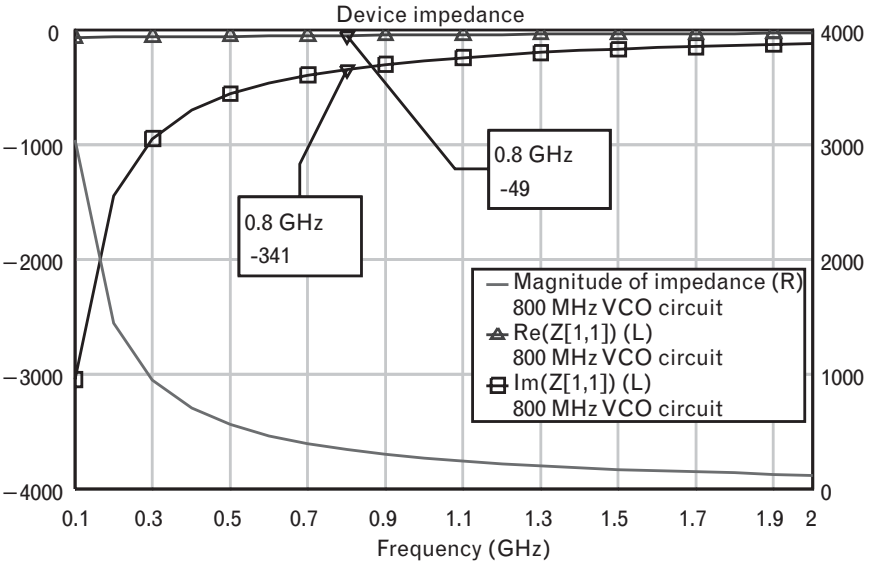


FIGURE 6.44
The impedance seen at the output port of the device in Figure 6.43, showing the magnitude and its real and imaginary parts.

831 MHz to 798 MHz. With the varactor capacitance set at 2.82 pF as shown, the oscillation frequency will be exactly 800 MHz. We can calculate the reactance slope with frequency using values 1 MHz either side of this oscillation frequency, by noting the reactance changes 45Ω over a 2-MHz frequency increment. We can then use (6.34) to estimate

$$Q_L = \frac{f_0}{2|R_D|}\left.\frac{dX}{df}\right|_{f_0} = \frac{800*10^6}{2(49)}\frac{45}{2*10^6} = 184 \tag{6.44}$$

This expression uses the oscillator model of Figure 6.14. The reactance slope in the above expression is set by the resonant circuit itself, whose X is dominant. The value of device resistance to use is difficult to estimate, but we have taken the small-signal value from Figure 6.44, which is its largest possible magnitude. As the coupling, or step-up, capacitor C_x is decreased, the reactance slope seen by the device at resonance increases, and the oscillator loaded Q also increases. Varying C_x is a very effective way of changing the relationship between the loaded Q and external Q, and thus trading off output power and phase noise.

6.1.4 Characterizing oscillator phase noise

If the oscillator used in a radio to create or to select channels is noisy, then in a transmitter the oscillator noise will spill into adjacent channels. In a receiver it will impair the quality of the desired channel, since reciprocal mixing can occur when a strong interfering signal mixes with phase noise down to IF, and falls on top of the IF resulting from a weaker desired signal.

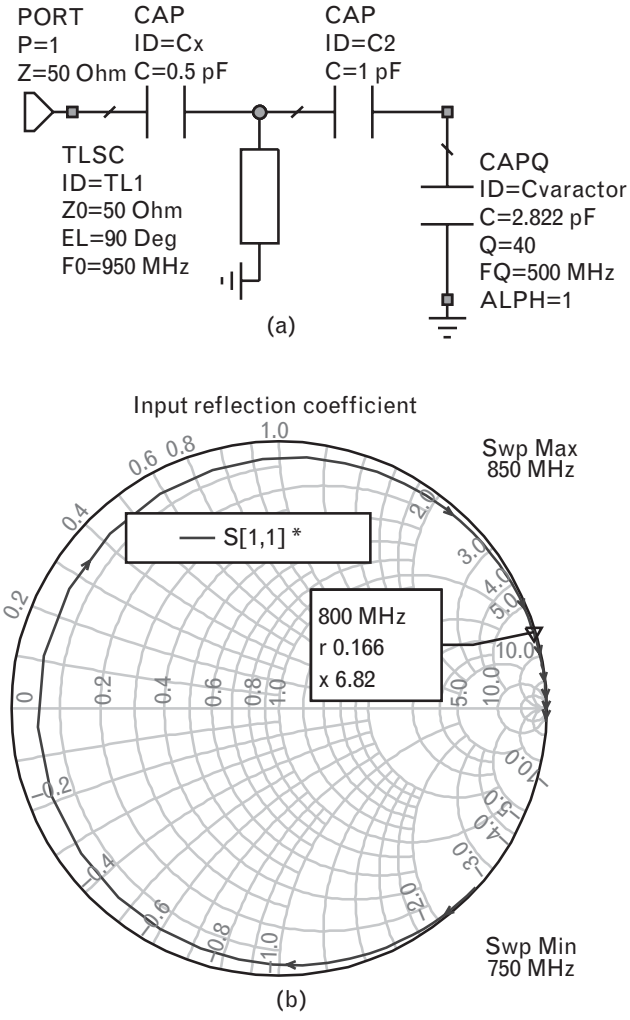


FIGURE 6.45
(a) The resonant circuit of the Clapp oscillator of Figure 6.42, loading the device of Figure 6.43. (b) The input reflection coefficient of the resonator circuit.

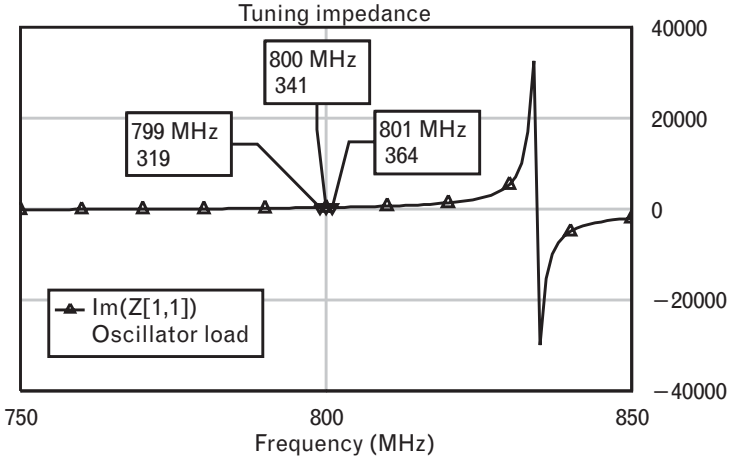


FIGURE 6.46
The reactance of the resonant load of the circuit of Figure 6.42 around the oscillation frequency

Leeson [14] first proposed the model for phase noise that we give next. His result is that the jitter or noise of an oscillator is inversely proportional to its power and inversely proportional to the square of the oscillator loaded Q, and that when comparing oscillators of different frequencies, the phase noise scales with the square of the operating frequency. We now set out to derive this result.

6.1.4.1 Signal modulation and an expression for phase noise

Any general modulated signal may be represented in the form

$$i(t) = A(t)\cos\left[\omega_C t + \phi(t)\right] \tag{6.45}$$

where the signal contains an amplitude modulated component represented by $A(t)$ and a phase modulated component represented by $\phi(t)$. The unmodulated carrier frequency is given by ω_C, and the instantaneous frequency is defined by the derivative of the total phase

$$\omega(t) = \frac{d}{dt}\left[\omega_C t + \phi(t)\right] = \omega_C + \frac{d\phi(t)}{dt} \tag{6.46}$$

It is helpful in (6.45) to break the signal down into its amplitude and phase modulated parts separately. If we consider the amplitude modulation to be sinusoidal modulation at some relatively slow modulation frequency ω_m, then (6.45) becomes

$$i(t) = \left(A + \Delta A\cos\omega_m t\right)\cos\left[\omega_C t\right] \tag{6.47}$$

A is the peak amplitude variation. The product term in (6.47) can be expanded to give

$$i(t) = A\cos\omega_C t + \frac{\Delta A}{2}\left[\cos(\omega_C + \omega_m)t + \cos(\omega_C - \omega_m)t\right] \tag{6.48}$$

This is a familiar result that simply states for sinusoidal amplitude modulation, the modulation sidebands are equal in amplitude $\Delta A/2$ and offset from the carrier frequency an amount equal to the modulation frequency ω_m. The single-sided spectral power density, which is the power in a sideband at a particular frequency component, is just

$$\left(\frac{\Delta A}{2A}\right)^2$$

relative to the power in the carrier.

We can derive a similar result for the phase modulation in (6.45) by assuming that the carrier is sinusoidally phase modulated at a frequency of ω_m with a small peak phase deviation $\Delta\phi$

$$i(t) = A\cos\left[\omega_C t + \Delta\phi \sin \omega_m t\right] \tag{6.49}$$

Expanding the cosine term that contains the sum of two arguments,

$$i(t) = A\cos\omega_C t\cos(\Delta\phi \sin \omega_m t) - A\sin\omega_C t\sin(\Delta\phi \sin \omega_m t) \tag{6.50}$$

Because the cosine of a small argument is approximately equal to one, and the sin of a small argument is approximately equal to that argument, we can further write

$$\begin{aligned} i(t) &\approx A\cos\omega_C t - A\sin\omega_C t(\Delta\phi \sin \omega_m t) \\ &= A\cos\omega_C t - \frac{A\Delta\phi}{2}\left[\cos(\omega_C + \omega_m)t - \cos(\omega_C - \omega_m)t\right] \end{aligned} \tag{6.51}$$

This result, for narrowband phase modulation, also has two sidebands spaced at offset frequencies $\pm\omega_m$ from the carrier ω_C. Just as for amplitude modulation, we can define the single-sided spectral density of the phase modulation as the ratio of modulation power in a 1-Hz sideband at offset frequency ω_m to the power in the carrier frequency, that is,

$$\left[\frac{\Delta\phi}{2}\right]^2$$

We can derive alternative expressions for the single-sided power spectral density from (6.46) by noting that

$$\omega = \omega_C + \Delta\phi\frac{d}{dt}(\sin\omega_m t) = \omega_C + \Delta\phi\omega_m\cos\omega_m t \tag{6.52}$$

so that $\Delta\omega = \Delta\phi\omega_m$ and $\Delta\phi = \Delta\omega/\omega_m = \Delta f/f_m$. These are all peak quantities since they are the coefficient of a sinusoid; in this case the equivalent *root-mean-square* (rms) terms equal the peak quantity divided by $\sqrt{2}$.

Now if the phase modulation is caused by some phase fluctuation in an active device, the *single-sided* power spectral-density ratio defined above is called the *phase noise*, and we can write a general expression for the phase noise $\mathcal{L}(f_m)$in a phase modulated signal as

$$\mathcal{L}(f_m) = \left(\frac{\Delta\phi}{2}\right)^2 = \left(\frac{\Delta f}{2f_m}\right)^2 = \left(\frac{\Delta\phi_{rms}}{\sqrt{2}}\right)^2 \tag{6.53}$$

In an oscillator, the phase noise can be read from a spectrum analyzer by comparing the power in a 1-Hz bandwidth at an offset frequency f_m to the power in the total oscillator signal. (Some corrections might need to be applied to account for the actual measurement bandwidth, and the type of detector, and we assume the spectrum analyzer itself has a noiseless LO.) This is a single-sideband measurement, and it also assumes that the AM noise of the oscillator is negligible compared with its phase noise.

The expressions above define the phase noise for any signal that encounters a phase fluctuation. However, these were derived by considering a signal with sinusoidal modulation, so that the two sidebands are correlated and equal (since they come from the same modulation source). More generally, we can define the total *power spectral density* (PSD) of the phase fluctuations $S_\theta(f_m)$ as

$$\begin{aligned} S_\theta(f_m) &= ave\left(\Delta\phi^2(f_m)\right) \\ &= \Delta\phi_{rms}{}^2(f_m) \end{aligned} \tag{6.54}$$

where we have first averaged the relative power in both sidebands (represented by the square of the peak phase deviation). By definition, this is just the rms value (i.e., the square root of the average value of the phase deviation squared). Thus, for instance, a 1° rms phase jitter corresponds to $10\log(1\star\pi/180)^2$, or –35 dBc signal-to-noise ratio between the carrier and the spur causing the jitter.

For a modulated signal as above, the power spectral density of the phase fluctuations will be 3 dB higher than the phase noise. This is reflected in the $\sqrt{2}$ in the rms term in (6.53), characteristic of the peak to root-mean-square ratio for sinusoidal modulation. Nor is the 3-dB difference unexpected, because the single-sideband power in each of the two sidebands is equal and adds directly to give the total power density at an offset frequency f_m.

However, depending on the measurement made, the measured power spectral density will not always be 3 dB higher than the phase noise defined by (6.53). If, for instance, the phase noise is measured by first downconverting the carrier to dc, the current components in (6.51) at $\omega_C \pm \omega_m$ will fold upon themselves after downconversion. This will double the *current* amplitude at the offset frequency ω_m compared with the original signal, thereby increasing the measured *power* spectral density 6 dB above the real phase noise derived from a single sideband. On the other hand, for the *uncorrelated* spectral components that are downconverted, such as the thermal noise, the rms values of power (rather than current) add when they are folded down to baseband, since the average in (6.54) is now that of a random quantity. Thus, the measured noise floor at baseband then increases by just 3 dB (rather than 6 dB) compared with the phase-noise floor of the oscillator. We

have neglected here the effect of any noise from the second oscillator needed to drive the downconverting mixer, which would increase the reading a further 3 dB if its phase noise contribution were equal to the oscillator being measured. During measurement all these effects are usually accounted for by applying a calibration factor (e.g., –6 dB) to the measured power spectral density of the downconverted signal, to correct the reading to a single-sideband phase noise equivalent about the carrier.

6.1.4.2 Signal modulation in a closed-loop system and oscillator phase noise

The expressions above are for a general signal in any open-loop system that might encounter phase modulation, such as in an amplifier. They can now be applied to an oscillator by using the general feedback model of Figure 6.1. In the model of phase noise that we will use [11, 14], let us assume that the forward loop gain $AL(s)$ is set equal to unity and the feedback circuit $H(s)$ is a resonator. We can do this without loss of generality if we assume that the forward gain has no effect on the phase modulation other than to introduce the phase fluctuation $\Delta\phi$ in the first place. Equation (6.2) then simplifies to

$$\frac{V_O(j\omega)}{V_I(j\omega)} = \frac{1}{1 - H(j\omega)} \tag{6.55}$$

For a series resonator, the input-output of the feedback loop represented by the transfer function $H(j\omega)$ will appear as in Figure 6.47.

Assuming simple voltage division across the feedback resonator, we can write the response of the feedback filter around its resonance as

$$\begin{aligned} H(j\omega) &= \frac{R}{R + j\omega L - \dfrac{1}{j\omega C}} = \frac{1}{1 + j\dfrac{\omega_0 L}{R}\left(\dfrac{\omega}{\omega_0} - \dfrac{\omega_0}{\omega}\right)} \\ &= \frac{1}{1 + \dfrac{jQ2\Delta\omega}{\omega_0}} \end{aligned} \tag{6.56}$$

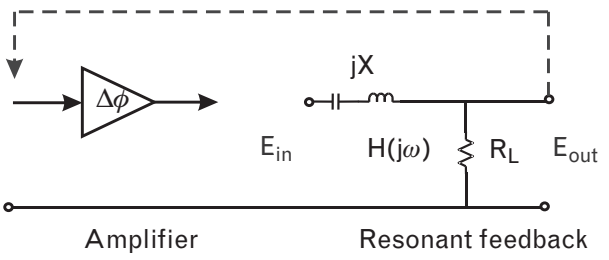


FIGURE 6.47
Resonator model for the closed-loop oscillator system.

using (6.10) and the fact that $\omega^2 - \omega_0^2 = (\omega - \omega_0)(\omega + \omega_0) \approx \Delta\omega \cdot 2\omega$ if the frequency is close to the resonant frequency so that $\omega \approx \omega_0$ and the difference is $\Delta\omega$. If $\omega = \omega_0$ exactly, then $H(j\omega_0) = 1$ as required by the Barkhausen criterion.

Thus, the output of the closed-loop oscillator system from (6.55) becomes

$$\begin{aligned}\frac{V_O(j\omega)}{V_I(j\omega)} &= \frac{1}{1 - \dfrac{1}{1 + \dfrac{jQ2\Delta\omega}{\omega_0}}} \\ &= \frac{1 + \dfrac{jQ2\Delta\omega}{\omega_0}}{\dfrac{jQ2\Delta\omega}{\omega_0}} \\ &= \frac{\omega_0}{jQ2\Delta\omega} + 1\end{aligned} \tag{6.57}$$

Therefore, from this expression, the equivalent power spectral density at the output of the closed-loop system is just the power spectral density of the input signal times the square of the closed-loop transfer function above, that is,

$$\begin{aligned}\mathcal{L}_{OSC}(\omega) &= \mathcal{L}_{IN}(\omega)\left|1 + \frac{\omega_0}{jQ2\Delta\omega}\right|^2 \\ &= \mathcal{L}_{IN}(\omega)\left[1 + \frac{\omega_0^2}{4Q^2\omega_m^2}\right]\end{aligned} \tag{6.58}$$

where $\mathcal{L}_{OSC}(\omega)$ is the phase noise at the output of the closed-loop system and $\mathcal{L}_{IN}(\omega)$ the phase noise introduced by the device itself. In (6.58) we have replaced the difference frequency $\Delta\omega$ by ω_m, since both are the offset frequency from resonance if we assume that the oscillation frequency is the resonant frequency of the feedback loop itself.

Furthermore, (6.58) is only valid for frequencies within the 3-dB bandwidth of the resonator, since we have simplified (6.56) by assuming small values of frequency offset. The 3-dB frequency of the resonator is given by

$$f_{3-dB} = \frac{f_0}{2Q_L} \tag{6.59}$$

where the Q is now taken to be Q_L, the oscillator loaded Q, since the load resistor in Figure 6.47 is that of the entire resonant system seen by the device. Outside the 3-dB frequency of the resonator, the system is essentially open-loop since the resonator then provides no feedback. The closed-loop phase noise will be just that of the open-loop system at large offset frequencies, usually the system noise floor itself.

Equation (6.58) is a fundamental expression describing the phase noise within a closed-loop system. The oscillator phase noise is just the phase noise of an open-loop signal, as might be described by (6.53), modified by a term that falls off as the inverse square of the offset frequency from the oscillation frequency. This expression is valid for as long as we measure within the 3-dB bandwidth of the resonator, where the final expression in (6.56) is valid. Outside this frequency range, the transfer function of the feedback filter $H(j\omega)$ is close to zero because either $j\omega L$ or $1/j\omega C$ dominate the denominator of the first expression in (6.56). Therefore, once we are at frequencies beyond the 3-dB bandwidth of the resonator, in (6.55) the closed-loop output voltage mirrors the input voltage and the phase noise at the output is just that at the input (i.e., the noise floor).

As a conclusion to our derivation on phase noise, we need to consider the form of the input phase noise in (6.58). At very low frequencies, below some corner or flicker frequency f_K, the phase noise contributed by a device rises above the noise floor inversely with frequency. This so-called $1/f$ noise is the phase noise introduced by the device itself, and it results from the creation of electron-hole pairs within the transistor due to the emptying of surface states, other impurities, or traps. The emptying and filling of these traps takes place with a very long time constant, so that as we approach dc the noise level from them can be well above the noise floor. For most devices, the corner frequency is in the kilohertz to megahertz frequency range. For Si BJTs and HBTs, it can be as low as 5 kHz. For Si MOSFETs, it is around 100 kHz, and for GaAs MESFETs it can be as high as 20 MHz. In an oscillator, this noise is upconverted about the carrier frequency, so it has a $1/f_m$ dependence, where again f_m is the offset frequency.

There will also be a baseline contribution to the phase noise arising from thermal noise itself. Its power spectral density is given by

$$P_{NOISE} = FkT \tag{6.60}$$

where the noise figure F accounts for the white noise added by the device. If the signal power is P_S, we can model the phase shift $\Delta\phi$ created by the thermal noise. Assume a phasor voltage of rms amplitude $\sqrt{P_S}$ rotating at the oscillation frequency f_0, and two superimposed noise voltage sidebands each of rms amplitude $\sqrt{FkT}$ rotating about its tip at offset frequencies f_m. This is shown in Figure 6.48.

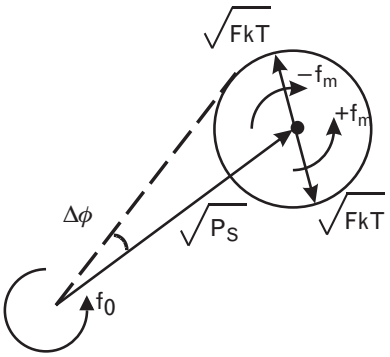


FIGURE 6.48
The additive effect of thermal noise on the signal, creating a phase error $\Delta\phi$.

Since the noise is random, *each* noise phasor produces a peak phase shift $\Delta\phi$ with amplitude given by

$$\Delta\phi = \sqrt{\frac{FkT}{Ps}} \tag{6.61}$$

or of rms amplitude

$$\Delta\phi_{rms} = \frac{1}{\sqrt{2}}\sqrt{\frac{FkT}{P_S}}$$

Since the powers from the two noise phasors are additive, the total rms phase fluctuation at f_m is given by

$$\Delta\phi_{rms\,(\,total\,)} = \sqrt{\frac{FkT}{P_s}} \tag{6.62}$$

We can now use (6.53) to calculate the single-sided power spectral-density, together with the $1/f$ frequency dependence, to obtain

$$\mathcal{L}_{IN}(f_m) = \frac{FkT}{2P_S}\left(1 + \frac{f_K}{f_m}\right) \tag{6.63}$$

In this expression, f_K is the radial corner frequency at which the effect of the $1/f$ noise has fallen to 3 dB above the noise floor, and the input phase noise is again measured as a function of the offset (or modulation) frequency f_m. It is a relative measurement because the entire signal vector in Figure 6.48 rotates at f_0 and the noise sidebands that create the phase error rotate with it at an incremental frequency f_m.

Substituting (6.63) into (6.58), we can write an expression for the output phase noise of an oscillator

$$\mathcal{L}_{OSC}(f_m) = \frac{FkT}{2P_S}\left(1+\frac{f_K}{f_m}\right)\left[1+\frac{f_0^2}{4Q_L^2 f_m^2}\right] \tag{6.64}$$

A typical plot of this expression is given in Figure 6.49. This plot shows the phase noise, measured in dBc/Hz. The single-sideband power is measured relative to the total carrier power, as defined by (6.53). The frequency axis is plotted as the logarithm of frequency and is measured as the offset frequency from the oscillating carrier (f_m in the above expressions). At large f_m, the phase noise of the oscillator is just the phase noise of the device itself as given by (6.53) and (6.62), that is,

$$\mathcal{L}(f_m) = \frac{FkT}{2P_S} \quad f_m >> f_K, f_{3-dB}$$

For example, if $F = 1$ and $P_S = -10$ dBm, this would be –167 dBc/Hz at room temperature where $kT = -174$ dBm/Hz. The astute reader will have noticed that the factor of 2 in the denominator of the equation above would appear to give only one-half the expected noise power ratio $FkT/2P_S$ far from the carrier. Remember however, (6.53) for $\mathcal{L}(f_m) = (\Delta\phi_{rms})^2/2$ is derived for the power in just one sideband *relative to the carrier*. The ratio of the total power in both sidebands to the carrier power, or the total power spectral density of the noise, is $(\Delta\phi_{rms})^2$ and double the power in a single sideband, assuming symmetry. This total noise power spectral-density, or FkT for thermal noise, is traditionally measured only for positive frequencies and assumes the mirror effect of negative frequencies is folded in. In practice, the noise floor in a radio will be greater than $FkTB$, where B is the preceding filter bandwidth, as it will include the noise from system components preceding the oscillator (and mixer).

The top diagram of Figure 6.49 shows the behavior of the spectral density of the device noise itself, given by (6.63). At frequencies less than

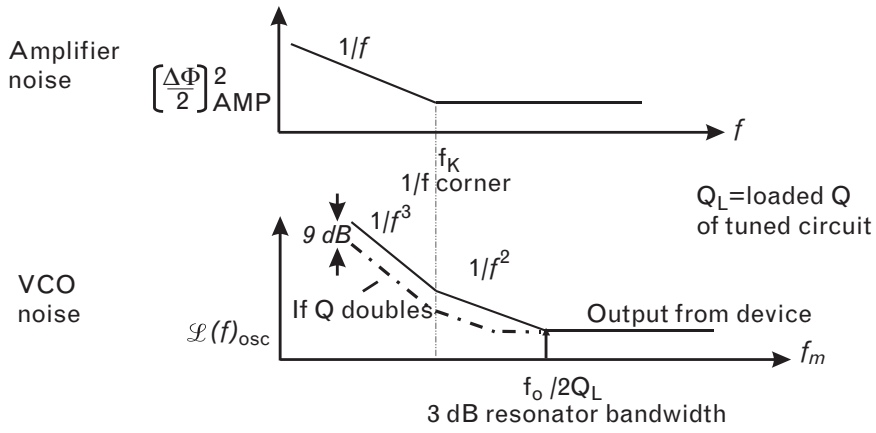


FIGURE 6.49
Typical phase noise plot for an oscillator showing characteristic frequency dependence.

the corner frequency, the noise rises inversely with frequency, or at 10 dB/decade. The phase noise of the closed-loop system is given by (6.64) and is shown in the bottom diagram of Figure 6.49. It depends on the position of the 3-dB resonator bandwidth relative to the device corner frequency. For low Q oscillators, with a resonator bandwidth higher than the device corner frequency, the close-in slope will be 30 dB/decade. However, the noise will then drop at 20 dB/decade at frequencies above the noise corner frequency until it reaches the noise floor. This is because the 3-dB resonator bandwidth is greater than the corner frequency and its effect varies as $1/f_m^2$. For high Q oscillators, the resonator bandwidth could be less than the device corner frequency. Its $1/f_m^2$ dependency would then result in an additional 20-dB/decade rise in the phase noise close-in to the carrier, to yield a total close-in slope of 30 dB/decade as before, but further out, the slope would only be the 10-dB/decade decay of the $1/f_m$ noise.

The oscillator power P_S is the total area under the (double-sided) power spectrum after subtracting the $FkT/2$ noise floor.

Phase noise is also expressible as the rms time jitter on the carrier signal. Expressing noise as jitter is sometimes useful in frequency multiplication or division, since it remains constant for any multiplication or division ratio. It can be calculated from the phase noise as

$$t(f_m) = \frac{\left(\mathcal{L}_{OSC}(f_m)\right)^{\frac{1}{2}}}{2\pi f_0} \tag{6.65}$$

6.1.4.3 Simulation and control of phase noise

The requirements to design an oscillator for minimum phase noise are apparent now from (6.64):

1. Maximize the loaded Q, by maximizing the resonator Q and coupling the resonator tightly to the oscillating device, and by minimizing the coupling of the load to the circuit. Similarly, saturation of the active device can also lower the loaded Q since the device losses will then add to those of the resonator. A 10-dB increase in loaded Q results in a 20-dB improvement in phase noise.
2. In a voltage controlled oscillator, maintain the Q_O of the resonator by avoiding forward bias on the varactor tuning diodes, limiting the signal swing across the tuning diodes to prevent heating and thermal effects. This can be achieved by placing the varactor circuit in the gate or base if possible.
3. Choose an active device with the lowest corner frequency and the lowest noise figure. MESFETs are notorious for impurities in the GaAs layer under the gate, and the resultant traps give them very

poor phase noise properties, reflected in their high corner frequency. Si BJTs, HBTs, and SiGe devices all have superior phase noise performance to MESFETs. In these devices, an unbypassed emitter resistor of 10Ω to 30Ω can improve the flicker noise by as much as 40 dB because of the negative feedback it provides to the device to reduce AM to PM conversion [11]. Of course, the selection of the device also needs to be made on the basis of its output power, the desired frequency of oscillation, as well as its flicker noise properties. As described in Chapter 5, the maximum power out of a device as an oscillator is when it is driven at its point of maximum power added efficiency, generally close to the 1-dB compressed output power. The oscillator output power is then the amplifier output power less the input drive power needed to sustain it. The maximum frequency range of the device is given by f_{MAX}, the frequency at which the maximum available gain drops to unity.

4. Maintain a high P_S/kT ratio. This term comes from the $\Delta\phi$ of the device itself. This phase perturbation can be minimized by using high impedance devices such as FETs, where the signal-to-noise ratio of the signal voltage relative to the equivalent noise voltage can be made very high [11]. The noise from the varactor diode resistance in the case of a varactor-tuned VCO can also become the dominant noise source. For good phase noise, the carrier signal effectively appearing across the varactor noise resistance should be maximized to maintain good signal-to-noise ratio at this point. By transforming the noise load resistance seen by the oscillating device to a lower value in the matching circuit, the power-to-noise ratio $V^2/4R_v$ across the varactor can be maximized, although at the expense of tuning bandwidth since the matching circuit will restrict the obtainable capacitance variation. However, there is a compromise with (2) above in order to avoid breakdown, saturation, or overheating effects in the varactor. These will all reduce the loaded Q.

5. Maintain a 90° crossing angle between the device line and the load line for the oscillator. It has been implicitly assumed in deriving (6.64) that the crossing angle is 90°, because the resonator model we have used in (6.56) assumes a series model with the series resonator and load added together to give the total resistance R. AM-PM conversion is minimized by choosing a 90° crossing angle between the device line (modeling the amplitude dependence) and load line (modeling the phase or frequency dependence). An example is given at the end of this chapter.

There are other factors that can minimize phase noise that are not apparent in this model and are generally determined experimentally. These

include minimizing *frequency pushing* by the gate or base voltage. Frequency pushing is a shift in the oscillation frequency caused by a change in the transistor bias voltage.

For example, when we considered device noise in an open-loop system and closed the loop to create oscillations, we implicitly assumed that phase noise is generally attributed to the transistor low-frequency or baseband noise that upconverts via a mixing process in the device into frequency fluctuations around the carrier signal. One mechanism for the noise and mixing in an FET oscillator is the modulation of its nonlinear gate-source capacitance caused by traps. Output fluctuations, for instance in the drain current generator, have much less impact on the phase noise. The upconversion process can then be modeled to first order by using a *pushing factor* on the input-referred baseband noise. The pushing factor is the oscillator frequency sensitivity to a change in the dc gate bias (measured in hertz/Vrms). It can be measured by modifying the gate bias voltage, or by superimposing a white noise source on the gate bias through the bias tee, and determining the frequency shift. Experimentally, the pushing factor can drop to nearly zero at a particular gate-bias point. Unfortunately, although the phase noise does have a minimum, it does not drop to near-zero, so that more complex models that rely on more than a single gate control voltage are necessary to simulate the upconversion of phase noise [15].

Other mechanisms for phase noise include downconversion of harmonic noise, and shot noise from forward conduction if the device is driven into saturation [8]. This occurs when the device load line is voltage-limited rather than current-limited (i.e., if the load resistance is too high). Using the two-port, open-loop design approach we described at the beginning of this chapter enables the choice of an appropriate load line for the open-loop amplifier to avoid voltage limiting and saturation. It also enables the small-signal open-loop gain to be kept relatively low to avoid deep compression at steady state. However, as we have seen, this is very dependent on the terminations that result when the loop is closed, and on s_{12} of the device. Nevertheless, knowledge of the load line can at least assist in keeping the steady-state compression relatively low, and in minimizing the nonlinear mixing effects that cause upconversion and downconversion.

Commercial CAD suites such as Microwave Office have recently introduced phase noise simulation in oscillators and require the input of just two model parameters to describe the $1/f$ noise of the device. For example, for bipolar transistors, the Gummel-Poon parameters *AF* (flicker noise exponent) and *KF* (flicker noise coefficient) are required. A harmonic balance analysis of the oscillator is performed to calculate the conversion matrix from baseband to the carrier frequency. The circuit's baseband noise is then upconverted using this matrix, and the resulting noise power spectral density can be plotted.

6.1.4.4 Phase-noise impact on system performance

When comparing phase noise between oscillators, the measurements need to be normalized to account for differing oscillation frequency and/or measurement offset frequency, using (6.64). For example, phase noise results of –160 dBc/Hz at 100-MHz carrier frequency quoted in [16] at 1-kHz offset frequency from the carrier frequency would scale to –140 dBc/Hz at 1 GHz and –120 dBc/Hz at 10 GHz. These can be compared with other measured state-of-the-art benchmarks of –133 dBc/Hz at 4.85 GHz [16] or –135 dBc/Hz at 1 GHz [17].

Typical system specifications for phase noise requirements in oscillators are:

- WCDMA –90 dBc/Hz at 10 kHz and –113 dBc/Hz at 100 kHz;
- GSM –111 dBc/Hz at 100 kHz and –143 dBc/Hz at 3 MHz;
- DECT –85 dBc/Hz at 100 kHz;
- Bluetooth –119 dBc/Hz at 3 MHz;
- Wireless LAN –116 dBc/Hz at 3 MHz.

Figure 6.50 shows one reason for the requirements on minimum phase noise in an oscillator. If a strong interfering signal is in the same channel as the desired signal, it will mix with an ideal LO to produce a downconverted interferer that exceeds the desired signal. In practice, it is more likely that a strong interfering signal will lie in an adjacent channel offset from the desired signal as shown. There, it can reciprocally mix with the LO to produce an offset downconverted signal, but the oscillator phase noise will also be linearly translated down as well. The noise may well be strong enough at an offset frequency corresponding to the downconverted position of the desired signal to substantially interfere with it. The minimum interferer level above the desired signal that produces a noise that is, say, 20 dB below the desired downconverted signal is one way to specify the requirement. Good detection of the desired signal in the presence of noise

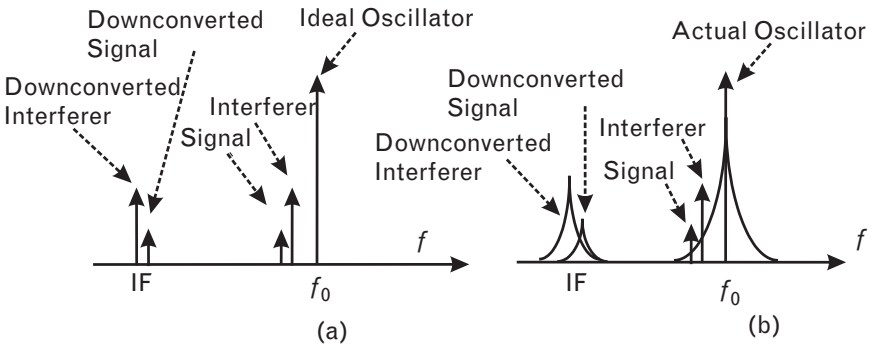


FIGURE 6.50
Reciprocal mixing from LO phase noise in a receiver.

from an interfering signal 70 dB stronger than it is considered a very good result. This effect is examined numerically for an example of a transceiver design in the final chapter of this book.

The same effect can arise when an interfering signal mixes with a spur from the oscillator to downconvert to the same IF as the desired signal. Typical systems require all spurious sidebands to be at least 70 dB down from the LO.

A second reason for controlling oscillator phase noise is to limit any smearing of the phase constellation of the downconverted signal. In a radio with a phase-modulated signal, it is the phase jitter (in degrees) that determines the degradation between adjacent allowed symbols in the signal constellation. For instance, symbols in the QPSK constellation are spaced 90° apart, but this will be corrupted an amount $\Delta\phi$ by the oscillator phase noise. The total phase jitter can be evaluated from (6.53) by integrating over the entire LO noise spectrum:

$$\left(\Delta\phi_{rms}\right)^2 = 2\int \mathcal{L}\left(f_m\right)df_m \tag{6.66}$$

where the upper integration limit is usually set by the channel or modulation bandwidth and the lower limit by the locking bandwidth of the system phase-lock loop.[8] The phase jitter produced by any spurious signals can also be calculated in a similar way from (6.54) and will add to the above result.

6.2 Oscillator design examples

6.2.1 45.455-MHz Colpitts crystal oscillator design

In this example we will design a 45.455-MHz Colpitts oscillator, using a crystal operating at its third overtone. The device selected is a BFS505 bipolar transistor from Philips Semiconductors operated with 3-V collector bias and 5-mA quiescent current. The topology used is that of Figure 6.38, in which the output power is coupled out of the collector of the device, loaded here directly with a 50-Ω resistor.

The input *S*-parameters of the device are plotted in Figure 6.51 up to 1,000 MHz. The first step is to derive a simplified device model for the input, to ensure consistency with the Colpitts design assumptions made in Section 6.1.3.2 and in Figure 6.33. At low frequencies, s_{11} lies on a line of constant conductance with shunt capacitance. At 50 MHz, the normalized

8. The bandwidth of the *phase-lock loop* (PLL) will typically be between 0.1% to 1% of the channel bandwidth, to avoid the PLL and the oscillator locking onto and tracking the carrier modulation rather than the carrier itself. The modulation frequency is usually just less than the channel bandwidth. Phase noise at frequencies lower than the PLL bandwidth will therefore be tracked and consequently eliminated.

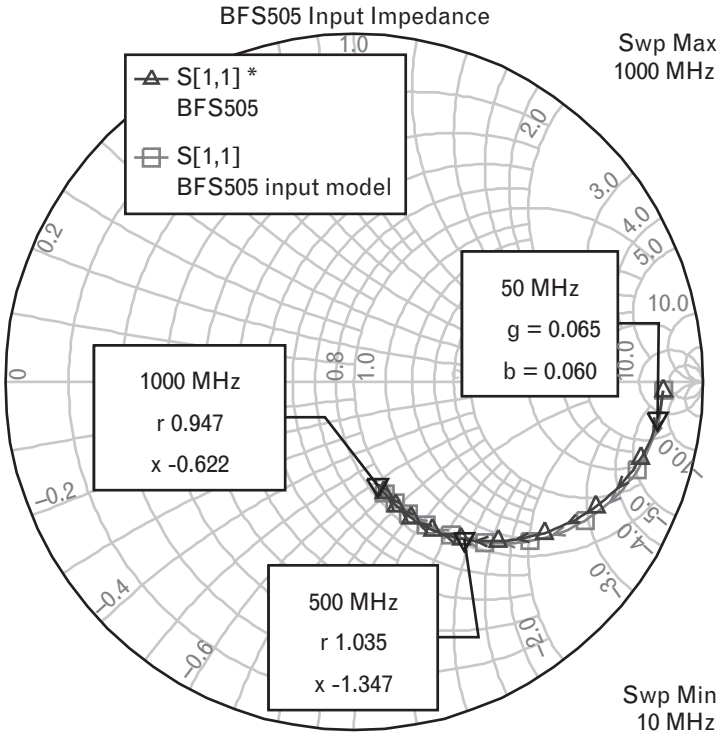


FIGURE 6.51
The s_{11} for the BFS505 transistor at 3V, 5 mA.

input conductance is about 0.065, or 770Ω, in parallel with a normalized susceptance of $j0.06$, or 4.2 pF. Thus, at low frequencies, to a first approximation, the base appears to be a parallel connection of $r_\pi = 770\Omega$, and $C_\pi = 4.2$ pF.

When biased at 5 mA, we estimate the emitter resistor of the transistor to be $r_E = 1/g_m = kT/qI_E$ or 26 mV/5 mA = 5.2Ω. The frequency at which the current gain becomes unity is therefore $f_T = 1/(2\pi C_\pi r_E) = 1/(2\pi \bullet 4.2 \star 10^{-12} \star 5.2) = 7.6$ GHz. This agrees well with the datasheet value of 9 GHz. The low frequency current gain is simply $h_{fe0} = r_\pi/r_E = 148$, so from Chapter 3, the 3-dB roll-off frequency can be calculated from the gain bandwidth product as $f_{3-dB} = f_T/h_{fe0} = 7.6 \star 10^9/148 = 51.4$ MHz. Above this frequency, the input capacitance starts to dominate over r_π, and the base junction begins to look increasingly capacitive. The series base resistance r_b then becomes a more important component of the input. This is indeed the case here, where from Figure 6.51, the input lies increasingly along a circle of constant resistance on the Smith chart as the frequency increases. At 1 GHz, we estimate the normalized resistance to be about 0.95, or 47.5Ω. The reactance change between 500 and 1,000 MHz is $+j.725$ (from $-j1.347$ to $-j0.622$) or $+j$ 36.2Ω. Since at this frequency the series reactance of the base model is simply that of the base inductance $j\omega L_b$ in series with the input capacitance $-j/\omega C_\pi$, we can calculate the difference in reactance between the two frequencies. Here, the reactance change is almost entirely due to the capacitance, and the inductance is negligible.

With this as a rough starting point, we can use a CAD tool to tune this simple model more accurately to the input data, and the equivalent input circuit given in Figure 6.52 results. The input impedance of the model is also plotted in Figure 6.51. The reactance value of the base capacitance, now optimized to 5 pF, is $-j700\Omega$ at 45 MHz. Knowing the base-loading on the Colpitts reactances X_1 and X_2, we can now commence the design itself.

Figure 6.53 shows the basic Colpitts oscillator topology. The crystal to be used is that described earlier in Section 6.1.3.4, and it has a third overtone resonance at 45.455 MHz. There the series resistance is 40Ω , so we estimate that the negative resistance measured from the base to ground in Figure 6.53 needs to be around -80Ω for steady-state oscillation. This should allow sufficient margin to still remain more negative than the crystal effective load resistance even after the transformation effects of the crystal parasitic capacitance.

With $g_m = 1/5.2$, and choosing $X_1 = X_2 = -50\Omega$ (70-pF capacitance at 45 MHz) and a series emitter resistor of 20Ω, (6.43) gives

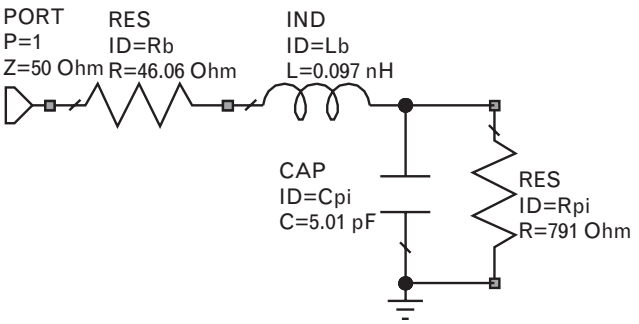


FIGURE 6.52 *The equivalent input circuit for the BFS505 transistor.*

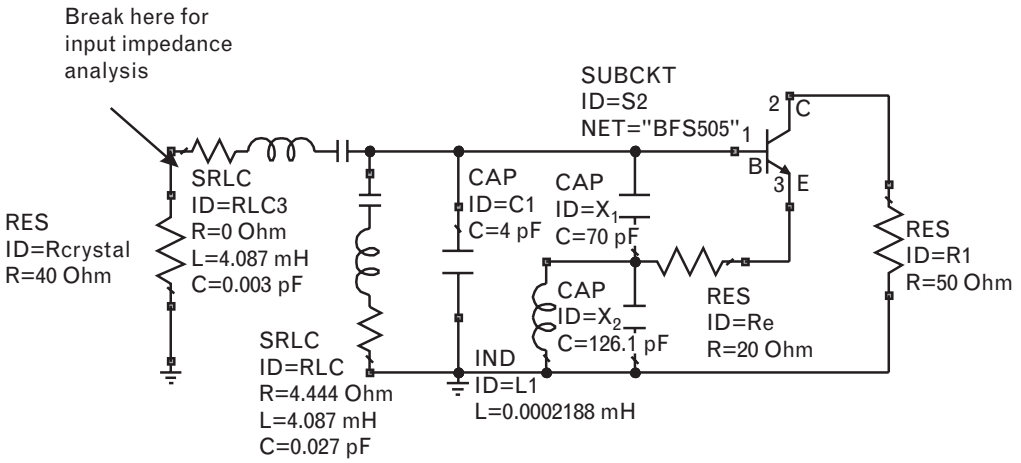


FIGURE 6.53 *The 45.455-MHz Colpitts oscillator.*

$$Z_{IN} = -\frac{X_1 X_2}{1/g_m + R_E} + j\left(X_1 + X_2\right) = -99 - j100 \qquad (6.67)$$

as the small-signal input impedance at the base of the Colpitts-connected transistor. We know that because the real part of the input impedance is dependent on g_m, the negative resistance will become less negative as the device compresses. However, we surmise that because the input resistance at the base is more than twice as negative as the crystal load, this is probably sufficient for startup of oscillations. We will need to verify this assumption.

We chose X_1 as a capacitive reactance of $-j50\Omega$ in order to completely swamp the internal reactance of the base capacitance ($-j700$) so that the base loading itself may be ignored (this was assumed in the derivation of the Colpitts equations). Similarly, we choose an external emitter resistor of value 20Ω so that it will dominate the internal transistor emitter resistance (5.2Ω). The value for X_2 is then chosen to yield an input resistance at the base that is sufficiently negative to sustain oscillations. Choosing X_2 to be equal to X_1 will maximize the equivalent series capacitance for a chosen negative resistance (proportional to the product of X_1 and X_2) and desensitize the oscillator to any changes in the input reactance of the transistor itself.

To oscillate at the desired third overtone of the crystal at 45 MHz rather than the fundamental at 15 MHz, X_2 must be inductive at the lower frequency and capacitive at the higher. We thus implement X_2 using a shunt L-C network, where the inductor and capacitor values are chosen so that the net reactance is $-j50$ at 45 MHz, and their resonant frequency is 30 MHz. This yields the values shown in Figure 6.53. At 15 MHz, the net reactance of X_2 will then be $+j27.8\Omega$ and the resulting input resistance at the base will be $+55\Omega$, ensuring the device cannot oscillate there.

6.2.1.1 Effect of the crystal parasitic capacitance

From (6.67), we calculated the expected small-signal Colpitts resistance at the base of the oscillating transistor to be $-99 - j100\Omega$, at the oscillation frequency. The parallel equivalent is a resistor of -201Ω in shunt with a capacitive reactance of $-j198\Omega$. This is the impedance imposed by the Colpitts circuit at the terminals of the crystal. However, the crystal resonance arm is also in parallel with the crystal parasitic capacitor of 4 pF. Although this parasitic capacitance is relatively small—its reactance at 45 MHz is only $-j875\Omega$—this adds in parallel with the Colpitts circuit reactance of $-j198$, yielding an equivalent net shunt reactance of $-j161$. The parallel resistance of -99Ω and this net shunt capacitance transform back to a series equivalent resistor of -79Ω in series with a capacitive reactance of $-j98$. Thus, the small-signal load seen by the crystal resonant arm has been transformed from -99Ω to a significantly smaller value of $-79\,\Omega$, totally due to the

parasitic shunt capacitance of the crystal. This transformation is shown in Figure 6.54, where the positive values of these resistances have been taken to avoid working in the extended Smith chart.

The second effect of the parasitic capacitance is to detune the oscillator. Whereas the effective reactive load produced by the circuit itself is $-j100\Omega$, equivalent to 35 pF or the two 70-pF capacitors in series at 45 MHz, this transforms to an effective load on the crystal series resonant arm

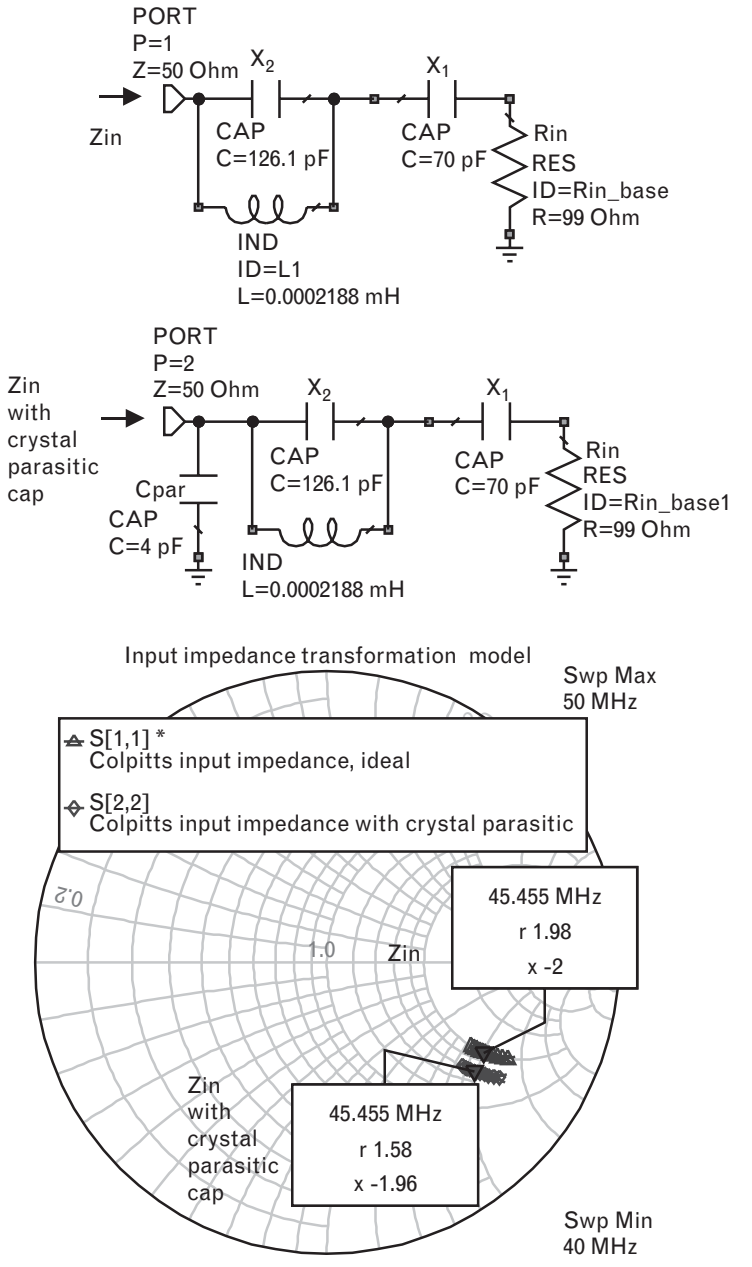


FIGURE 6.54
The expected effective impedance of the Colpitts device and the series resonant load at the crystal terminals, with and without the 4-pF crystal shunt capacitance. (Positive values of resistance are used instead of the actual negative values.)

of $-j98$, as shown in Figure 6.54. At the frequency of oscillation, therefore, the crystal will operate in its inductive region with a reactance of $+j98$ (instead of $+j100$). From (6.29) for a series resonant circuit,

$$\Delta f = \frac{\Delta X}{4\pi L} \tag{6.68}$$

so the series resonant arm will shift above its resonant frequency by an amount $98/4\pi\star4.087\star10^{-3} = 1{,}908$ Hz higher in frequency, in the crystal inductive region.

The resulting component values providing the third-overtone resonance, the negative resistance, and output load for the Colpitts oscillator can be clearly identified in Figure 6.53. Since the oscillator is a closed-loop system, we break the system for analysis at a convenient point, in this case at the crystal series resistor. At steady-state oscillation, we expect the impedance looking into this port to be equal and opposite to the load impedance, which is just 40Ω.

The simulation of Figure 6.55 was performed with measured *S*-parameters for the BFS505 at 3V, 5-mA bias. It shows a range of negative resistance values from above 30 MHz to more than 100 MHz, but only at 45.45682 MHz does the net reactance looking in through the crystal

Frequency (MHz)	Re(Z[1,1])	Im(Z[1,1])	|S[1,1]|
10	124.6	−5.049e+006	1
14.8	135.1	−3.205e+006	1
20	183.5	−2.139e+006	1
30	260.5	−9.994e+005	1
40	−109.7	−2.994e+005	1
45.455	−67.89	−93.87	1.577
45.4551	−67.89	−88.74	1.63
45.4552	−67.89	−83.6	1.69
45.4553	−67.89	−78.47	1.76
45.4554	−67.89	−73.33	1.839
45.4555	−67.88	−68.19	1.932
45.4556	−67.88	−63.06	2.04
45.4557	−67.88	−57.92	2.167
45.4558	−67.88	−52.79	2.317
45.4559	−67.88	−47.65	2.498
45.456	−67.88	−42.52	2.717
45.4561	−67.88	−37.38	2.984
45.4562	−67.88	−32.24	3.315
45.4563	−67.88	−27.11	3.725
45.4564	−67.88	−21.97	4.233
45.4565	−67.88	−16.84	4.848
45.4566	−67.88	−11.7	5.544
45.4567	−67.88	−6.567	6.199
45.4568	−67.88	−1.431	6.573
45.45682	−67.88	−0.4042	6.592
45.4569	−67.88	3.704	6.46
45.457	−67.88	8.84	5.927
50	−50.01	2.227e+005	1
60	−29.93	6.563e+005	1
70	−20.27	1.039e+006	1
80	−14.76	1.391e+006	1
90	−11.28	1.721e+006	1
100	−8.924	2.037e+006	1

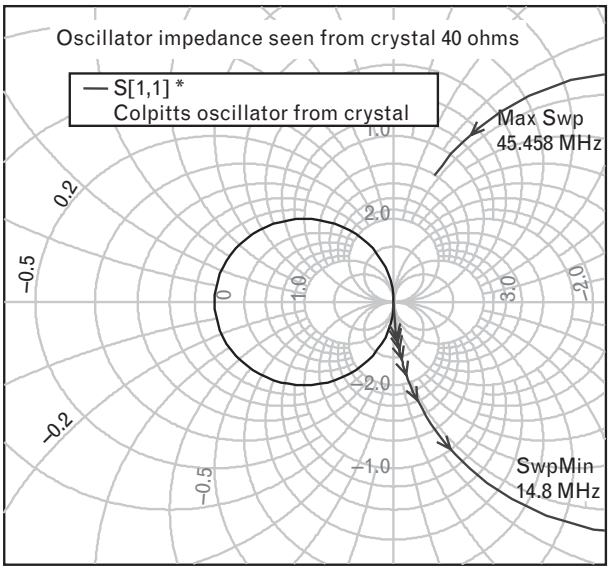


FIGURE 6.55 *The impedance of the Colpitts oscillator and common-base amplifier, looking in at the resistor of the crystal resonator.*

motional arm equal zero. This, therefore, is the frequency of oscillation, which is 1,820 Hz higher than the crystal resonant frequency. The slight discrepancy from the 1,908 Hz calculated arises because the reactance slope calculated in (6.29) is accurate only at the crystal series resonance, and it will change as we move away from it. At this frequency offset, the crystal presents an inductance of $+j98$ to compensate for the net capacitance of the shunt parasitic, X1, and X2, which amounts to $-j98$. Further away from resonance, the crystal capacitance and inductance present huge reactances that prevent oscillation at other frequencies.

The simulated small-signal value of device impedance at the crystal reference plane is -68Ω, compared with our derived value of -79Ω, excellent agreement in view of the assumptions that were made in reaching it. When large-signal *S*-parameters are used, with reduced values of s_{21} for the transistor to simulate the effect of device compression, the input resistance reduces to -60Ω. Since this is still more negative than the crystal resistance of 40Ω, we can be confident that oscillations will start up.

6.2.2 Design of a 3.7- to 4.2-GHz voltage-controlled oscillator

The Colpitts topology oscillator and variants of it are widely used at frequencies below about 1 GHz. At higher frequencies, the smaller g_m of the transistor, our inability to neglect the base loading of the transistor, and the transit time through it all make the Colpitts recipe less reliable. In this section, we will design an oscillator without any a priori assumptions, to tune across the 3.7- to 4.2-GHz frequency band, using a bipolar transistor, the BFR360F from Infineon Technologies. This low-cost device has a transition frequency of 14 GHz, good noise figure (0.95 dB), and a very low $1/f$ noise-corner frequency of 15 kHz. Tuning will be accomplished using a varactor at the base, where the voltage and current swings are the lowest. We need the oscillator to tune with an available tuning voltage between 0 and 5V, and require a minimum output power of +10 dBm.

6.2.2.1 Creating large-signal *S*-parameters for quasi-linear modeling

We will first use familiar quasi-linear modeling techniques to understand the expected device behavior before attempting full nonlinear modeling. To do this, we characterize the device by its large-signal *S*-parameters. This allows us to simulate the effect of gain compression as the oscillatory signal builds up. We can generate a variety of different *S*-parameter files, each corresponding to a different level of gain compression at 4 GHz, by driving the Gummel-Poon model for the BFR360F in a 50-Ω circuit in a nonlinear simulator. We justify this step only by later using the model for complete nonlinear simulation, once we have a good first-draft circuit.

We choose a bias voltage of 3V with 50-mA dc collector current, since the class-A 1-dB compressed output power is then at most 3*50/2 = 75 mW or about 18 dBm. The oscillator output power will be somewhat less; however, 10 dBm should be achievable. Table 6.1 lists the small-signal common-emitter *S*-parameters for the BFR360F at this bias point.

Table 6.2 lists *S*-parameters for the same device, but with a larger drive power at the input and output that result in about 3-dB compression of s_{21} at 4 GHz. Although reasonably accurate, this approach suffers from the fact that over a broad bandwidth in a 50-Ω system without any matching, the *input* compression point of a transistor occurs at increasing input power levels with frequency, since its gain drops. For example, a device driven into 1-dB compression at 4 GHz will already be about 7 dB into compression with the same *input* power at 2 GHz, assuming its gain at 2 GHz is 6 dB higher than at 4 GHz. This reflects the fact that the *output* compression point is principally determined by the device voltage and current swings along the load line, which are relatively invariant with frequency. Thus, for the same input power level, the compression at lower frequencies is much more severe, while at higher frequencies the transistor is barely compressed at all.

It can be seen that our earlier assumption that the magnitude of s_{21} is the first parameter to be affected by increased power is not a bad one, with its angle less sensitive to drive. However, especially for a bipolar transistor, s_{11} also begins to change rapidly as the device compresses, and strictly speaking, its change should also be characterized if we are to retain respectable modeling accuracy.

Table 6.1 Small-Signal Common-Emitter *S*-Parameters for the BFR360 (Bias Conditions: 3V, 50 mA)

F (GHz)	s_{11}	Ang	s_{21}	Ang	s_{12}	Ang	s_{22}	Ang
2.00	0.82	–165.01	6.35	95.89	0.04	30.58	0.21	–93.23
3.00	0.82	–172.73	4.27	88.08	0.05	32.86	0.18	–103.85
4.00	0.82	–177.45	3.21	82.41	0.05	35.39	0.18	–110.43
5.00	0.82	179.05	2.57	77.62	0.06	37.18	0.19	–115.22
6.00	0.82	176.17	2.15	73.28	0.07	38.13	0.20	–119.18
7.00	0.82	173.65	1.84	69.23	0.07	38.37	0.21	–122.74
8.00	0.82	171.36	1.61	65.37	0.08	38.03	0.23	–126.11
9.00	0.82	169.23	1.43	61.66	0.09	37.24	0.24	–129.39
10.00	0.82	167.21	1.29	58.06	0.09	36.11	0.26	–132.61
11.00	0.82	165.29	1.17	54.56	0.10	34.71	0.28	–135.79
12.00	0.82	163.43	1.08	51.15	0.11	33.12	0.29	–138.95

Table 6.2 Large-Signal Common-Emitter *S*-Parameters for the BFR360 (3V, 50 mA) Measured with Constant Input Power Applied First at the Input and Then at the Output Port, Where the Device Is About 3 dB Compressed at 4 GHz

F (GHz)	s_{11}	Ang	s_{21}	Ang	s_{12}	Ang	s_{22}	Ang
2.00	0.52	–126.17	2.26	109.50	0.04	30.49	0.21	–96.66
3.00	0.62	–143.10	2.19	99.40	0.05	32.92	0.18	–107.43
4.00	0.70	–156.43	2.11	90.32	0.05	35.51	0.18	–113.88
5.00	0.75	–166.98	1.98	82.68	0.06	37.31	0.19	–118.43
6.00	0.78	–175.09	1.84	76.27	0.07	38.27	0.20	–122.12
7.00	0.80	178.59	1.67	70.66	0.07	38.49	0.21	–125.43
8.00	0.81	173.43	1.53	65.74	0.08	38.13	0.23	–128.56
9.00	0.82	169.53	1.39	61.48	0.09	37.33	0.24	–131.62
10.00	0.82	167.23	1.26	57.82	0.09	36.18	0.26	–134.65
11.00	0.82	165.25	1.15	54.32	0.10	34.78	0.28	–137.67
12.00	0.82	163.37	1.06	50.92	0.11	33.17	0.30	–140.68

6.2.2.2 Loading the device in the unstable region

The next step is to examine the stability circles of the device. Because the required tuning bandwidth is less than 20%, we can do this at a single frequency in the middle of the band, using the small-signal and 3-dB compressed *S*-parameters to set the limits of the unstable region as the device compresses.

The device is relatively stable in its normal common-emitter configuration, but when mounted with common-collector, there are regions of instability within the Smith chart. Our intent in designing an oscillator in this frequency range will be to intentionally load the input (base) and output (emitter) of the device, now mounted common-collector, in the unstable regions. From Figure 6.56, we can deduce that the base requires an inductive load and the emitter a capacitive load to be unstable, corresponding to the top and bottom areas of the Smith chart, respectively. We also see that because the stability circles move towards the edge as the device gain is reduced, the Q of these load terminations must be high enough to maintain instability at steady state when the device will be compressed.

We first load the emitter with a capacitance and examine the impedance seen from the base. For now, the capacitance value should be tuned to ensure that around 4 GHz the input resistance of the device is negative across the band. Because of this negative resistance, the reflection coefficient at the base will be greater than one, and we will need to work on the

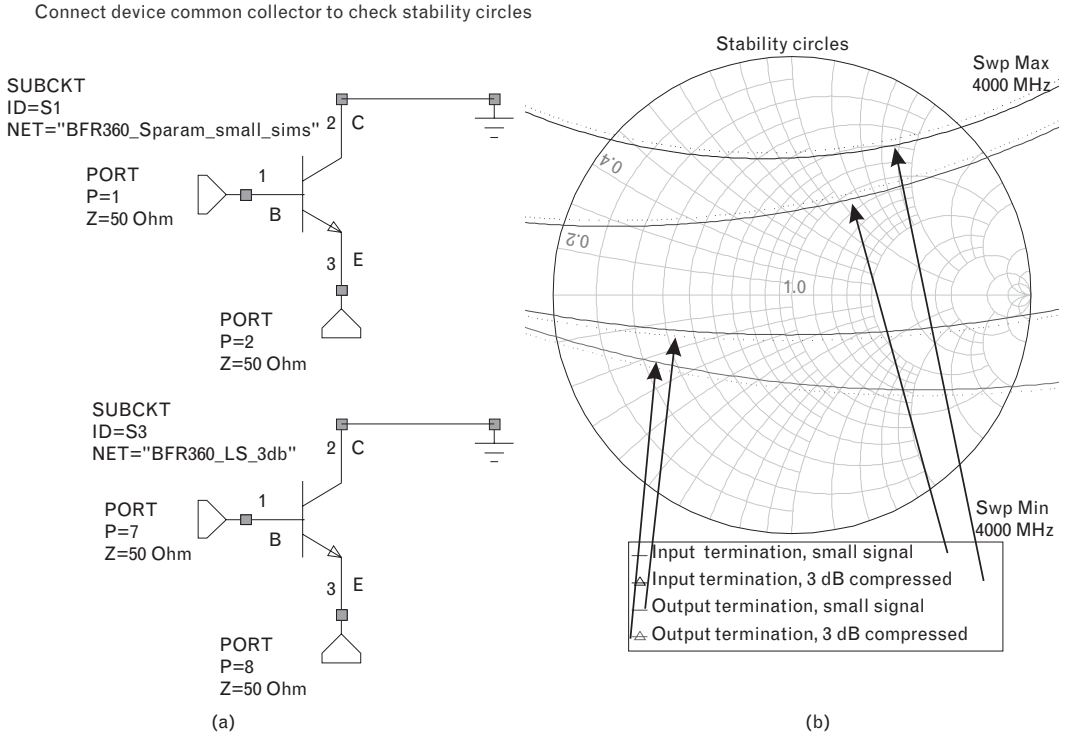


FIGURE 6.56 *(a) Simulation of the BFR360 in common-collector configuration at 4 GHz, using large-signal S-parameters to model small-signal and 3-dB compression. (b) The input and output stability circles for the configuration shown. The unstable region is on the same side as the dotted line.*

extended Smith chart. To avoid working in such unfamiliar territory, we will define a pseudo reflection coefficient $\Gamma(A)$ by inverting the actual reflection coefficient Γ_D. As shown in Volume I, Chapter 2, this is equivalent to changing the sign of the impedance Z_D.

Thus, if we can plot the inverse reflection coefficient of the device as it goes from small-signal to large-signal, we have plotted $Z(A)$, the quantity required for the oscillator impedance plots discussed in Section 6.1.2.4. Of course, linear simulators plot impedance as a function of frequency, not drive level, so that several simulations will be required using the different large-signal *S*-parameters to obtain the device impedance as a function of frequency, with compression level as the parameter. By choosing a constant frequency on each of the various simulations, and joining these impedances, the locus of $Z(A)$ is obtained.

Figure 6.57 shows such a locus at 4,000 MHz with $C = 1.2$ pF. The locus $Z(A)$ is shown in the direction of increasing drive A, corresponding to increasing compression. Its normalized reactance is fairly constant around $x = 1.3$, but its normalized resistance changes from $r = 1.14$ to $r = 0.63$ as the device compresses. This corresponds to the device resistance

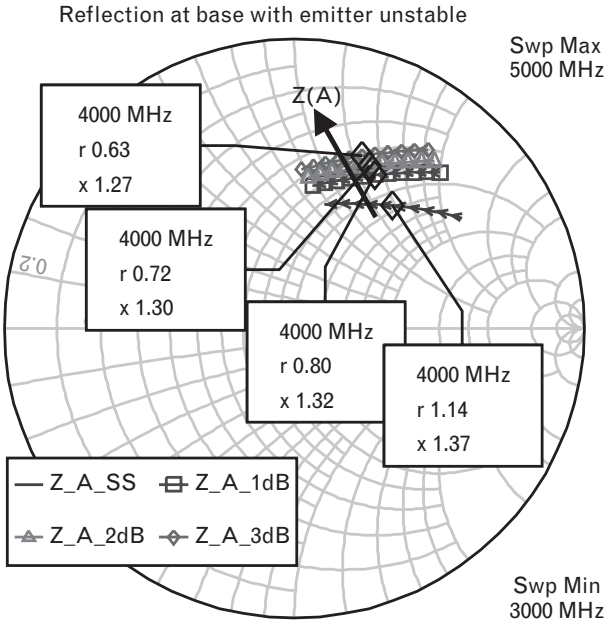


FIGURE 6.57
The inverse of the reflection coefficient looking into the base, simulated between 3,000 and 5,000 MHz, when the emitter is terminated with C = 1.2 pF. Simulations for small-signal, 1-, 2-, and 3-dB compressed S-parameters are shown, with Z(A) at 4,000 MHz the impedance joining the markers.

changing from –57Ω to –31.5Ω. We can deduce from this variation that the base needs to be terminated in an equal and opposite impedance [i.e., $Z(A)$ as shown], at the desired frequency of oscillation and compression level. Seen from the base, the transistor appears as a negative resistance device, and so requires a series R-L termination there. This will create a curve $Z_L(f)$ that lies on a circle of constant resistance with frequency, and create a crossing angle with $Z(A)$ close to the desired 90°. This will provide the most stable oscillator operating point and minimum phase noise, since any AM/PM conversion—between amplitude noise in the tuning reactance to frequency fluctuation—is minimized by keeping the intersection point tightly defined.

Correspondingly, we next load the base with an inductance and examine the nature of the impedance looking into the emitter. This intentionally loads the input of the transistor in its unstable region. To start with, the base inductance is set to a value that forces the resistance looking into the emitter to be negative around 4 GHz. We find that an inductance of 3.5 nH is a good start. $Z(A)$ is again plotted from a series of linear simulations using each of the large-signal S-parameter files and by inverting the simulated reflection coefficient at the emitter.

Figure 6.58 shows that the locus of $Z(A)$ now lies on a curve of nearly constant susceptance. While the normalized susceptance is around $b = 0.4$, the conductance changes from $g = 1.02$ at small-signal to $g = 0.50$ at large-signal levels, corresponding to a device resistance of –49Ω and –100Ω, respectively. The device resistance becomes more negative as the drive is increased. Thus, a series resistance is totally inappropriate to model the emitter behavior. Instead, we need to work in terms of admittance $Y_L(f)$

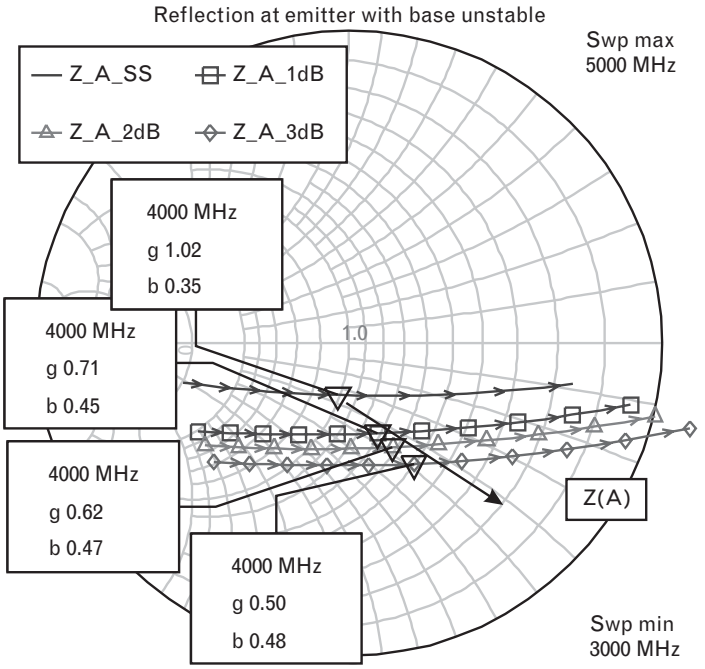


FIGURE 6.58
The inverse of the reflection coefficient looking into the emitter, when the base is terminated with L = 3.5 nH. Simulations for small-signal, 1-, 2-, and 3-dB compression are shown, with Z(A) the curve joining the markers at 4,000 MHz.

and $Y(A)$. $Y(A)$ moves along a line of nearly constant susceptance and decreasing conductance as the drive is increased, so $Y_L(f)$ needs to have a locus of constant conductance with frequency. The device and load curves will then cross at around 90° for minimum phase noise. We conclude that the load impedance at the emitter needs to be a capacitance in shunt with the load resistor.

6.2.2.3 Optimizing the desired termination impedances

Now that the nature of the base and emitter termination impedances is known, we optimize each in turn. Although described as a sequential procedure below, in practice some iteration between the separate steps will be necessary to obtain the desired tuning range.

We have determined that the base load impedance must be a series R-L circuit. The resistance is set initially to 1Ω to model the varactor loss, and the inductor is optimized so that the reflection coefficient looking into the emitter terminal Γ_D is maximized ($|\Gamma(A)|$ is minimized). This formalizes the earlier procedure where the inductor was set to 3.5 nH on the basis of manual tuning. The varactor reactance for now will be included in the total reactance modeled by this inductor, and broken out separately later.

Figure 6.59 shows the optimized base termination, and the resulting variation of the inverse of Γ_D with frequency. This simulation was performed at the 1-dB compression point so the transistor will operate close to its point of maximum power-added efficiency, thus peak oscillator output power. With an inductor of 4 nH, we have maximized the negative

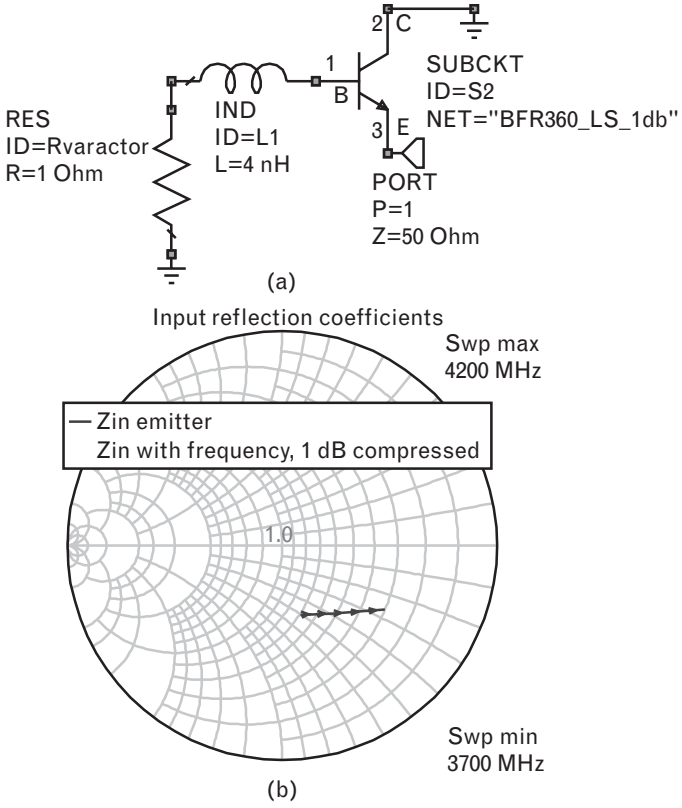


FIGURE 6.59
(a) Optimized base circuit for best reflection coefficient looking into the emitter, and (b) the resulting emitter response versus frequency. The inverse of the reflection coefficient is plotted so the negative of the emitter impedance or conductance is shown.

conductance looking into the device across the desired tuning range. The real and imaginary parts of admittance looking into the emitter, as well as the reflection coefficient, are tabulated in Table 6.3.

We now repeat the analysis used to create Figure 6.59 with the other large-signal *S*-parameter files. Figure 6.60 shows the emitter responses at

TABLE 6.3 THE NORMALIZED ADMITTANCE AND REFLECTION COEFFICIENT MAGNITUDE, $|\Gamma|$, LOOKING INTO THE EMITTER OF FIGURE 6.59 ACROSS FREQUENCY

F (MHz)	RE(Y) = G	IM(Y) = B	$\lvert\Gamma\rvert$
3,600	–0.01576	–0.01141	3.0846
3,700	–0.01387	–0.01006	2.9994
3,800	–0.01209	–0.00886	2.8032
3,900	–0.01043	–0.0078	2.5441
4,000	–0.00887	–0.00686	2.2698
4,100	–0.0074	–0.00601	2.0095
4,200	–0.00602	–0.00525	1.7774

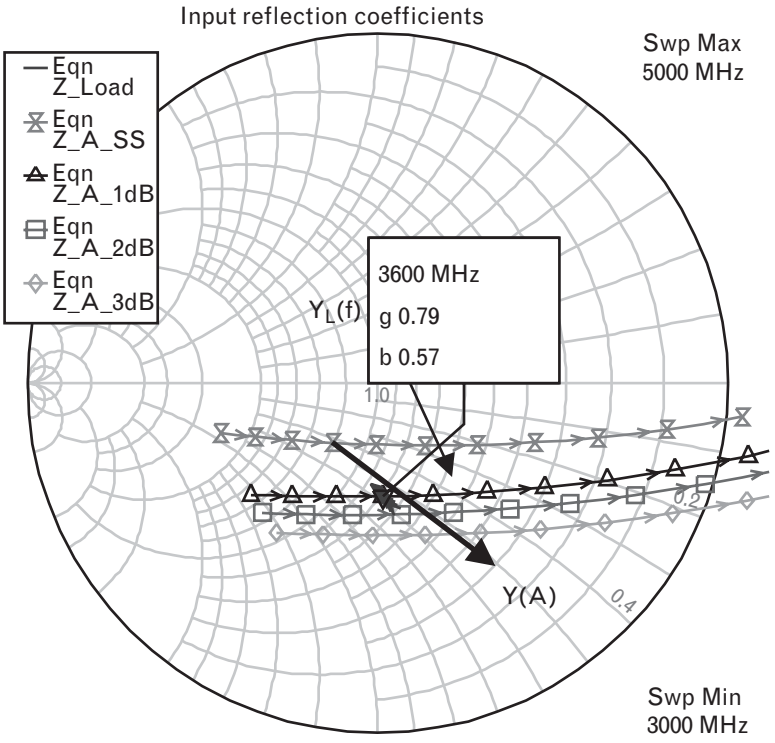


FIGURE 6.60
The device terminated with the optimized base circuit, showing the locus of Y(A) and Y_L with frequency, seen looking into the emitter.

different compression levels. From these, we can construct the locus of $Y(A)$. For reasons that will shortly become apparent, we do this first at 3,600 MHz.

Next, we design the load impedance to be placed on the emitter so that it is equal to $Y(A)$ at this frequency. This load needs to be matched to satisfy $Y_L(f) = Y(A)$ at 3,600 MHz *and* at a reasonable compression level to ensure sufficient output power from the device. We do this with the device compressed 1-dB. The desired normalized admittance value, $Y_L(f_0 = 3{,}600$ MHz$) = 0.79 + j0.57$, can be synthesized by a circuit of the form of Figure 6.61(a). Starting with the 50–Ω load that we will place at the emitter port, a series-C, shunt-C matching network will achieve the necessary transformation to $Y(A)$.

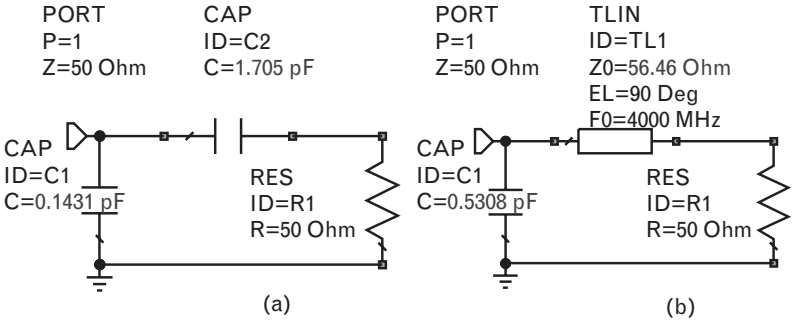


FIGURE 6.61
(a) One possible emitter load admittance satisfying $Y_L = Y(A)$ at 3,600 MHz. (b) An alternative implementation that provides better phase noise.

Although the device will indeed oscillate with this load at the correct frequency and amplitude, the load is suboptimal for phase noise, because its frequency variation, shown as $Y_L(f)$ in Figure 6.60, does not have a 90° crossing angle with $Y(A)$. In fact, it is barely visible in the figure because it lies almost parallel to $Y(A)$. Instead, the circuit of Figure 6.61(b) is a better implementation, because it provides the necessary shunt R-C termination for the emitter discussed in the previous section. The quarter-wave transmission line of characteristic impedance 56Ω transforms the 50-Ω resistor to a higher value, and the shunt capacitor then rotates the resistance along a circle of constant conductance until it satisfies $Y_L(f_0) = Y(A)$. Any variation with frequency is then along the circle of constant conductance, which is at 90° to the device line, which as noted earlier lies along an arc of almost constant susceptance. The variation of this load admittance $Y_L(f)$ with frequency is shown in Figure 6.62 together with the reproduced emitter admittance $Y(A)$.

Although the topology of the possible load circuit can be designed on the Smith chart and its values chosen quite close to $Y(A)$, an *exact* match should be obtained through optimization. If $Y_L(f_0) = Y(A)$, then $\Gamma_L(f_0)\Gamma_D = 1$. This latter product can be defined as an output variable calculated from a linear simulation. We can then optimize the load circuit in the simulator so

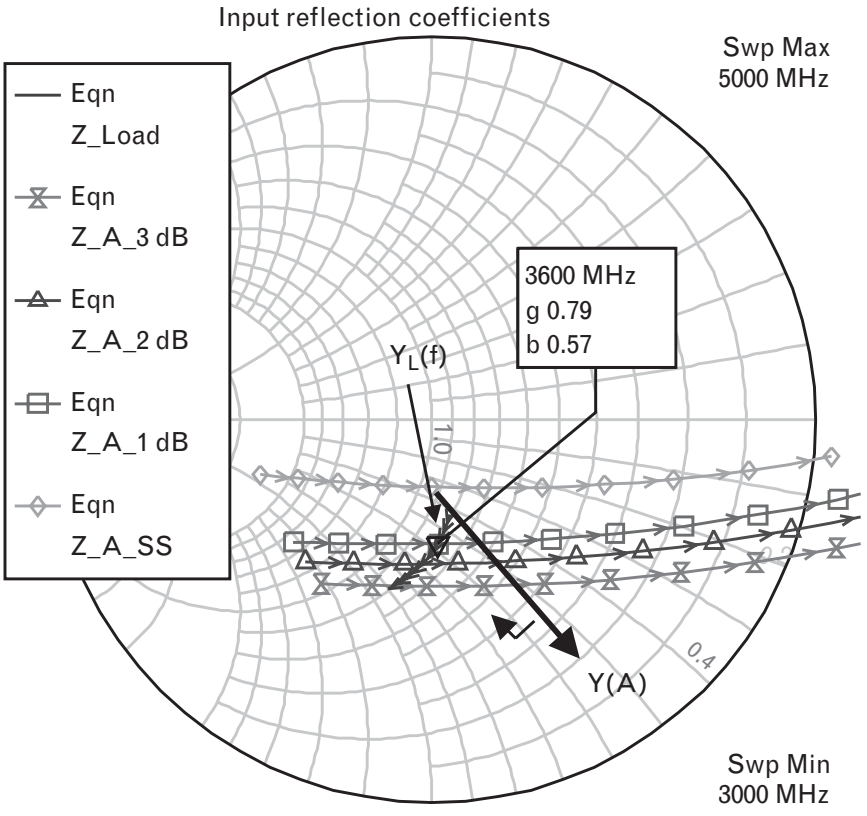


FIGURE 6.62
Variation of the emitter load admittance of Figure 6.61(b) with frequency, and the device admittance of Figure 6.59(a) with drive at 3,600 MHz. A 90° crossing angle has now been achieved.

that at the specified frequency, and using the desired *S*-parameter set, this variable is forced *exactly* equal to one. This ensures that when the loop is closed and the load impedance is connected to the device, the input admittance of the load is exactly equal and opposite to the input impedance of the device.

This circuit now meets the conditions for oscillation at 3,600 MHz and has embedding terminations at the base and emitter designed for 90° crossing angles to minimize the phase noise. The impedance seen at the base, when the 1-Ω varactor resistance is lifted from ground as in Figure 6.63, is shown in Table 6.4.

As expected, Table 6.4 shows that the impedance at 3,600 MHz looking into the device is zero, since the port was created by lifting the varactor resistor from ground and looking into the base through it. The reason for optimizing the design at 3,600 MHz should now be apparent, since only at frequencies above it is the resistance looking into the base negative. This encompasses the entire tuning range and allows for the possibility of oscillations across it. Had we instead optimized the load circuit at 4,000 MHz in the previous step, the base resistance would have been positive over those parts of the tuning range below 4,000 MHz and the circuit would not have oscillated there.

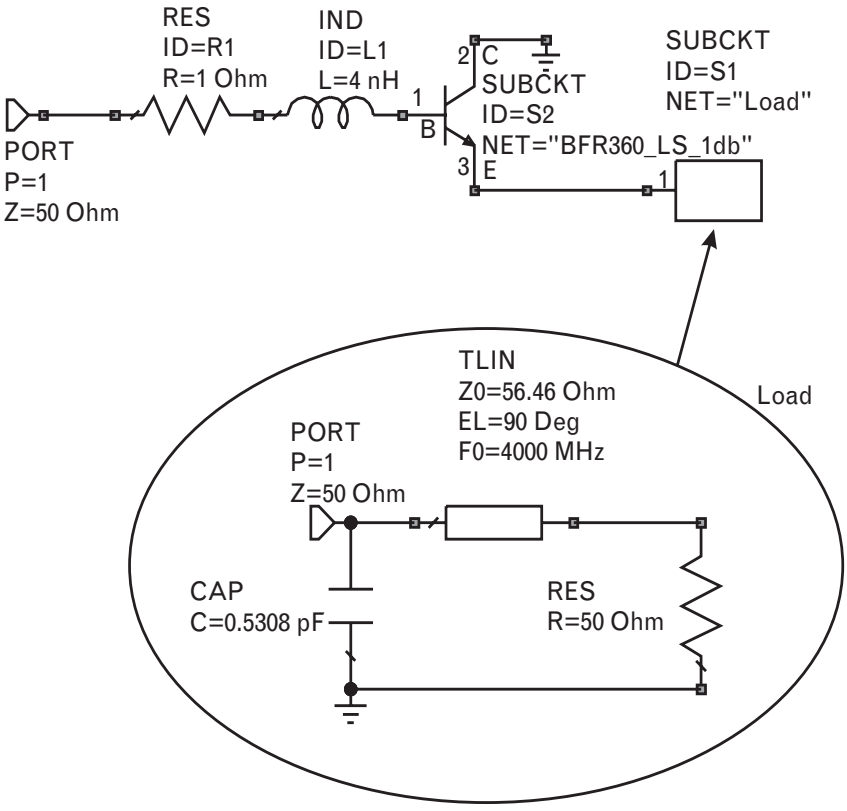


FIGURE 6.63
The oscillator circuit with optimized emitter load termination, used to examine the impedance seen at the base through the varactor resistor.

Table 6.4 The Input Impedance of the Oscillator of Figure 6.63, Looking into the Base Through the Varactor Resistor and Input Circuit

Frequency (MHz)	Resistance (ohms)	Reactance (ohms)
3,000	8.59	–28.60
3,100	6.92	–23.80
3,200	5.36	–19.00
3,300	3.88	–14.20
3,400	2.50	–9.44
3,500	1.21	–4.71
3,600	0.00	0.00
3,700	–1.13	4.69
3,800	–2.17	9.36
3,900	–3.14	14.00
4,000	–4.03	18.60
4,100	–4.84	23.20
4,200	–5.57	27.80
4,300	–6.24	32.30
4,400	–6.84	36.80
4,500	–7.37	41.20
4,600	–7.85	45.70
4,700	–8.26	50.00
4,800	–8.61	54.40
4,900	–8.91	58.70
5,000	–9.16	62.90

6.2.2.4 Achieving the tuning of the oscillator

Figure 6.64 is a plot of the resistance and reactance at the base corresponding to Table 6.4. It shows that the oscillator can be tuned to 3,700 MHz by adding a series capacitance at the base with reactance of $-j4.7\Omega$, of $-j18.6\Omega$ at 4,000 MHz, and of $-j27.8\Omega$ at 4,200 MHz.

Some of this additional reactance is added through the variable capacitive reactance of the varactor diode, whose value is a function of the voltage V across it. The capacitance of a varactor diode is of the form

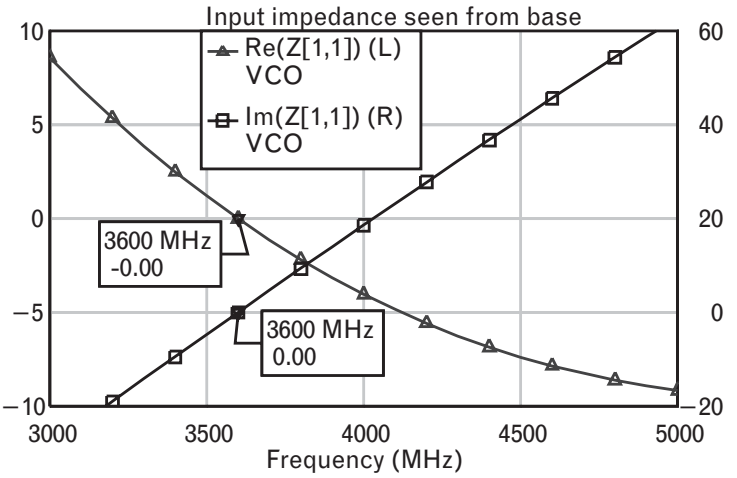


FIGURE 6.64
Total resistance and reactance looking into the base through the input termination and varactor resistor.

$$C_J(V) = \frac{C_J(0)}{(1 + V/V_S)^{\gamma}} \tag{6.69}$$

where $C_J(0)$ is the capacitance with no applied voltage. The reverse-bias voltage V is normalized to the semiconductor contact potential V_S which is 0.6 for silicon and 1.1 for gallium arsenide. The exponent γ will vary from between 0.3 up to 2, depending on the diode doping profile. For a step p-n junction, the exponent is 0.5 and the capacitance will vary inversely with the square root of the applied voltage; its reactance will increase as the square root of the voltage. At the other extreme, with an exponent approaching 2, the varactor is a hyperabrupt diode. By differentiating (6.69) and noting that the resonant frequency of oscillation is $1/\sqrt{LC}$, one can show that the oscillation frequency will (in theory) vary linearly with voltage. The series resistance of a varactor diode must not be neglected in modeling, since it can be significant. In our example, we use a MA46450 GaAs varactor chip from M/A-COM, Inc., which has V_S = 1.1V and γ = 1.0.

In order to be able to tune to frequencies between 3.7 and 4.2 GHz in Figure 6.64, the entire reactance curve needs to be shifted "down" so that at the desired frequency of oscillation the total reactance shifts to zero. A varactor diode may not be able to provide the necessary reactance variation, so it may also be necessary to add some fixed inductance in the series, and in some cases series capacitance, to adjust the slope (hence, Q) of this reactance curve across the desired tuning band. The choice of $C_J(0)$ and the allowable voltage variation in (6.69) determine whether any inductance needs to be added, and this can best be determined by simulation and checking the tuning bandwidth. Adding fixed inductance increases the capacitive reactance that must be added to compensate, but reduces the percentage variation it requires across the band. Since additional inductance increases the reactance slope, it will also increase the Q of the VCO.

Using a varactor with $C_j(0) = 1.5$ pF gives a good compromise for tuning bandwidth. Using a larger varactor capacitance requires less added series inductance but does not give as much reactance variation with varactor voltage, so reduces the tuning range. Using a smaller value requires a larger fixed series inductance to compensate for the increased added (capacitive) reactance. Although this increases the Q of the input circuit, as the varactor becomes smaller its losses will increase and its parasitics, which are fixed, become a larger proportion of the total capacitance, again reducing the total reactance variation.

The required ratio of maximum to minimum tuning capacitance may be calculated using the expression

$$\frac{C_{MAX}}{C_{MIN}} = \left(\frac{f_{MAX}}{f_{MIN}}\right)^2 + \frac{C_L}{C_{MIN}}\left(\left(\frac{f_{MAX}}{f_{MIN}}\right)^2 - 1\right) \tag{6.70}$$

where C_L is the residual or load capacitance of the oscillator itself (if nonzero), and of any fixed capacitance in the circuit appearing across the varactor terminals [8]. This had to be originally tuned out by the (4-nH) series inductance we added at the base. Its effect is to limit the potential bandwidth over which tuning can be achieved. The residual capacitance can be measured by looking into the base of the oscillator device at which we eventually add the varactor tuning circuit, and converting the reactance seen into an equivalent capacitance. In the case of Figure 6.63, with the inductor and varactor removed, the reactance at the base is $-j81\Omega$ at 4,000 MHz, equivalent to an input capacitance of 0.5 pF. Using (6.70) with f_{MIN} = 3,700 MHz and f_{MAX} = 4,500 MHz to cover the tuning range, and selecting C_{MIN} = 0.27 pF as the lowest realizable varactor capacitance, we calculate a capitance ratio of 2.37 and C_{MAX} = 0.64 pF. This capacitance range corresponds to tuning the 1.5 pF (zero-bias capacitance) varactor between a tuning voltage of 5V and 1.5V, almost identical to that we will simulate. The expression is useful for calculating other combinations of varactor capacitance and residual load capacitance that could achieve the desired tuning range.

The complete oscillator circuit is shown in Figure 6.65, where the base circuit is kept as an open port for analysis but grounded in actual operation. The varactor capacitance is now included, with additional fixed inductance to compensate for the excess reactance. The impedance looking into the input port from ground is simulated in Figure 6.66, which shows the reactance looking in when the varactor voltage is just under 3V. Since there is no net reactance at 4,000 MHz and we know from our earlier analysis that the net resistance is negative there, this corresponds to the frequency of operation.

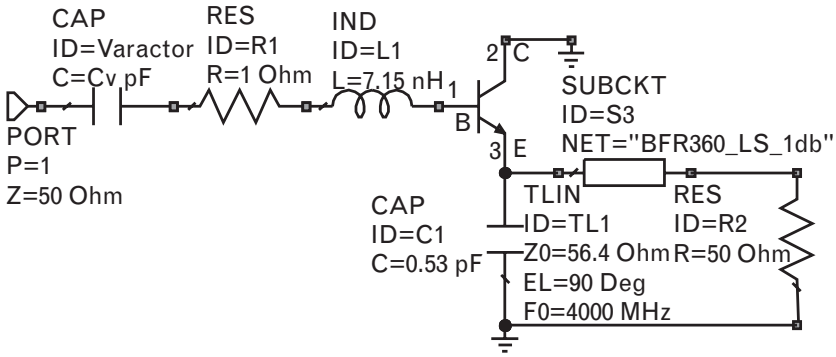


FIGURE 6.65
The completed VCO circuit. The input port is shorted to ground in operation but left open here for impedance simulation.

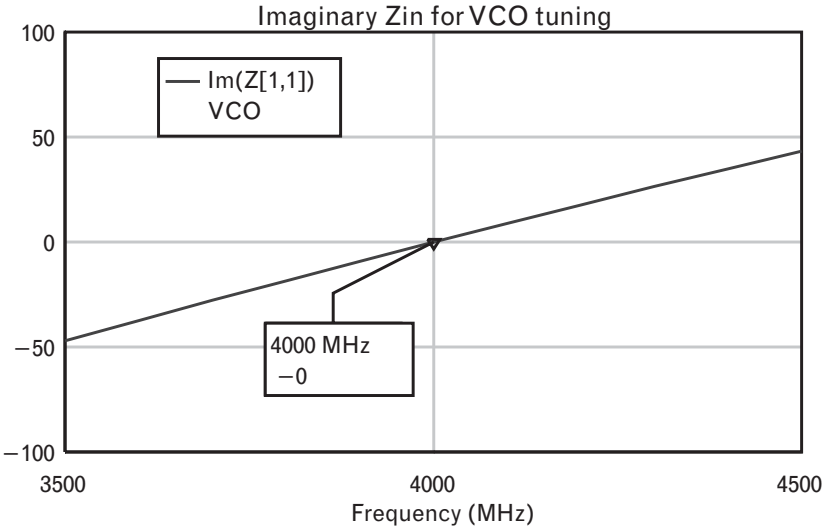


FIGURE 6.66
Input reactance of the oscillator of Figure 6.65 for a varactor voltage of 2.95V.

By repeating the simulation for a variety of different varactor voltages, we can look for the frequency at which there is zero net reactance looking into the input. Since the input is ultimately shorted, this gives the frequency at which the resistance and reactance of the device (the entire circuit of Figure 6.65) is equal and opposite to the reactance of the load (i.e., zero when the device port is shorted to ground). The tuning curve of Figure 6.67 results.

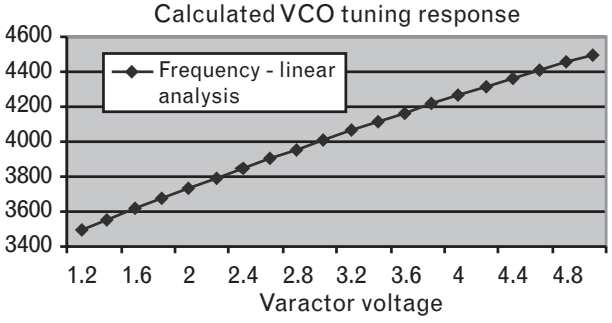


FIGURE 6.67
The tuning curve of the VCO of Figure 6.65.

6.2.2.5 Analysis of the entire oscillator

The quasi-linear analysis we have performed now requires multiple simulations, first across the different sets of *S*-parameters corresponding to drive (or compression), and second for different tuning voltages. The simulations for tuning voltages of 1.9V, 2.95V, and 5V are shown in Figures 6.68 through 6.70, respectively.

As in Section 6.2.2.3, we are now looking in at the port created when the emitter is split from its load, similar to the configuration of Figures 6.59 and 6.60. The load on the base is the series inductor and varactor, with the bottom end of the varactor diode again grounded. The admittance of the load with frequency does not change between Figures 6.68 to 6.70, but the device admittance seen at the emitter port shifts significantly with varactor voltage. The change in operating point is apparent from the figures, with the intersection setting the operating conditions: the frequency from $Y_L(f)$ and the degree of device compression from $Y(A)$. As the varactor voltage is increased, the degree of compression increases slightly, since the operating point shifts from the 1-dB compressed curve at 3,700 MHz to 3-dB compressed at 4,500 MHz. This indicates reasonably constant output power across the range. The crossing angle remains at 90° and indicates best phase noise for the given Q of the circuit.

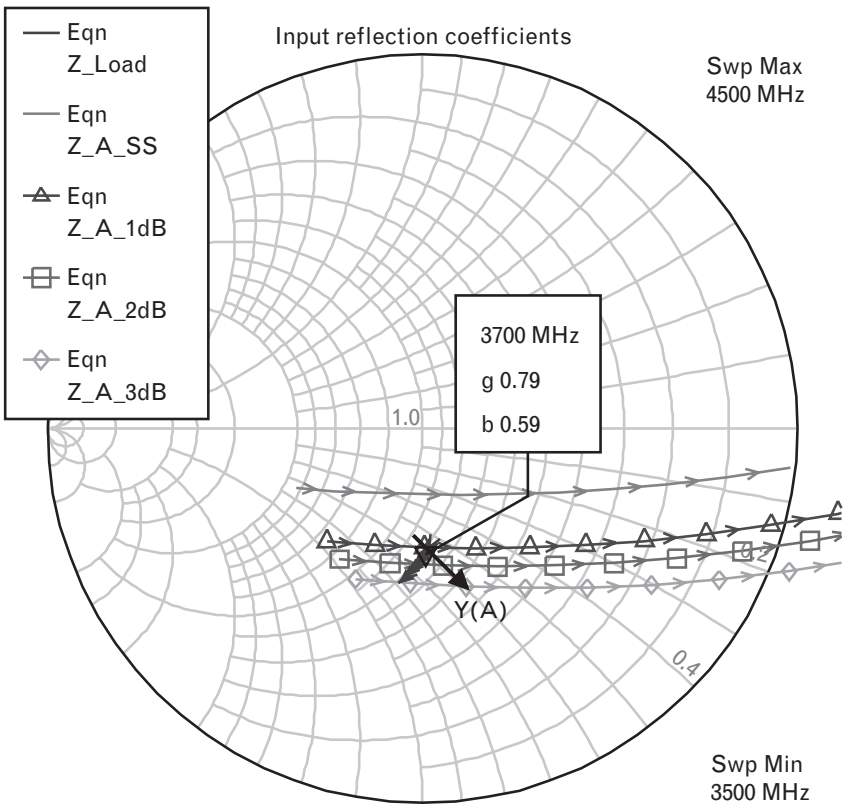


FIGURE 6.68
The device line looking into the emitter as a function of compression, and the load line of the emitter load as a function of frequency. Varactor voltage = 1.9V, frequency = 3,700 MHz.

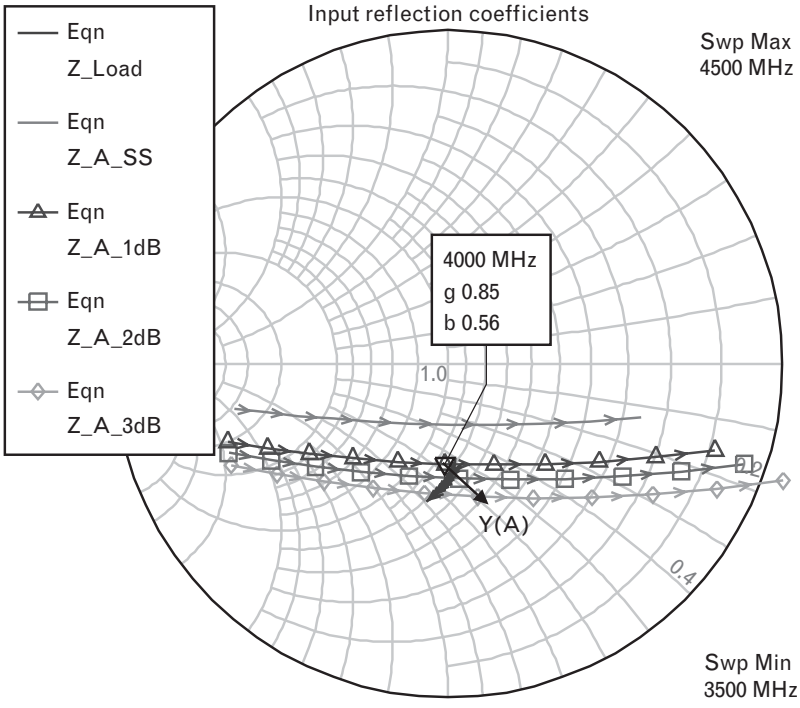


FIGURE 6.69
The device line looking into the emitter as a function of compression, and the load line of the emitter load as a function of frequency. Varactor voltage = 2.95V, frequency = 4,000 MHz.

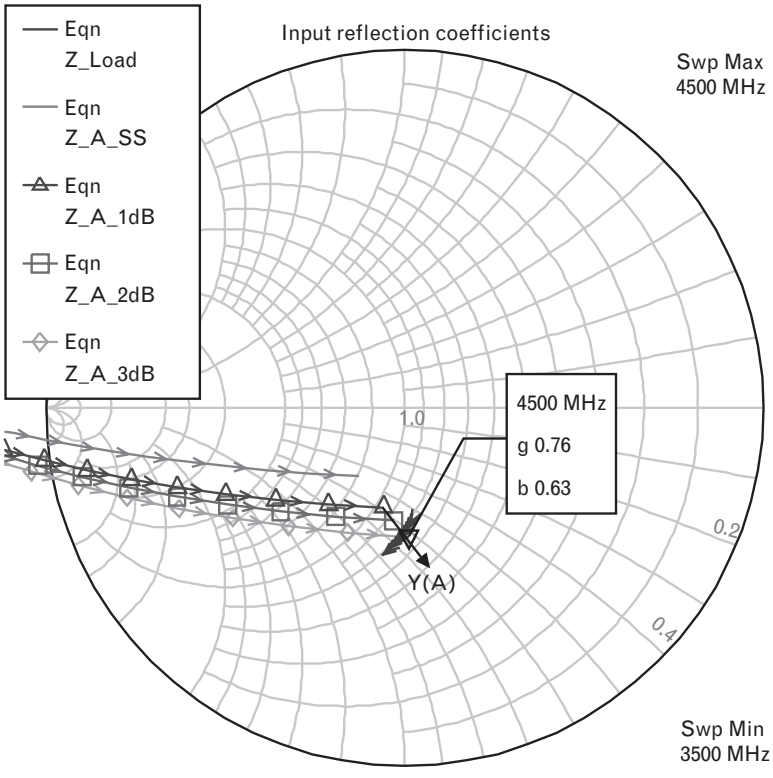


FIGURE 6.70
The device line looking into the emitter as a function of compression, and the load line of the emitter load as a function of frequency. Varactor voltage = 5V, frequency = 4,500 MHz.

6.2.2.6 Nonlinear analysis of the oscillator

The final step to validate the design is to return to the original Gummel-Poon model and quantitatively analyze the circuit we have just constructed. This can be done in a harmonic-balance simulator by using an oscillator probe, which is inserted into a suitable point between the device and the resonator in a closed-loop circuit. Initially, it applies an RF voltage at an estimated frequency and sweeps that frequency and voltage for values at which the probe supplies no current to the circuit. As described in Chapter 4, the probe has no effect on the circuit when this condition is achieved, since it supplies no power, and that solution point corresponds to the point at which the circuit is autonomous (i.e., oscillates without an applied source). The complete simulated circuit is shown in Figure 6.71. We will continue to model the varactor as a tunable capacitor for simplicity; it could also be modeled as a reverse-biased diode so the effect of the output RF voltage swing on the average capacitance could also be modeled.

The device is configured in common-collector by RF grounding the collector through the bias supply and providing an RF choke at the emitter for a dc ground. The simulations converged to a final solution once the bias

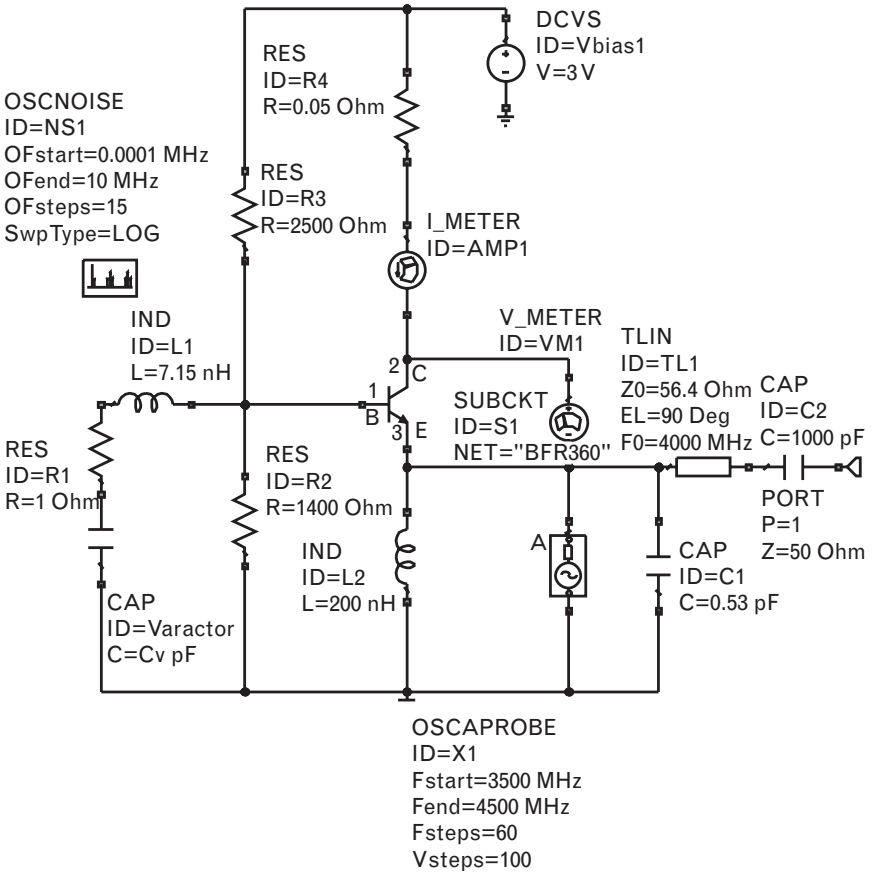


FIGURE 6.71
The VCO schematic for nonlinear analysis. The RF circuit is identical to that of Figure 6.65.

circuit was adjusted so it could source sufficient base current to support the oscillator output swing. The resistive divider required slight adjustment from that used in the initial *S*-parameter analysis in order to achieve that, even though the resulting steady-state dc collector current was approximately the same as for the linear simulations (65 mA instead of 50 mA).

A typical output spectrum is shown in Figure 6.72. There, the varactor voltage was set to 2.95V, and a fundamental output power of over 18 dBm was obtained at 4,025 MHz. The second harmonic is only 17 dB below this, so the transistor is relatively nonlinear. The output power is approximately equal to the 1-dB compression point of the device with the given 3-V bias conditions.

The dynamic load line of the VCO transistor is shown in Figure 6.73. The features that provide the limiting action of the oscillator are readily apparent, since the peak-to-peak voltage swing of 6V and peak-to-peak current swing of nearly 120 mA would be just as indicative of those in a healthy power amplifier: the device drives itself to the point of forward conduction and into cutoff at alternate ends of the cycle.

The tuning curve of Figure 6.74 results from sweeping the varactor voltage.

The agreement of the frequency versus varactor voltage characteristic with that derived from the linear simulations is quite remarkable, especially above 4,000 MHz where the difference is only a few megahertz. As the bottom end of the tuning range is approached, the oscillations start to die out, and the simulator can find no solution for an oscillation frequency when the varactor voltage is less than 1.4V. This is also indicated by the curve of output power, which starts to fall off rapidly at low varactor

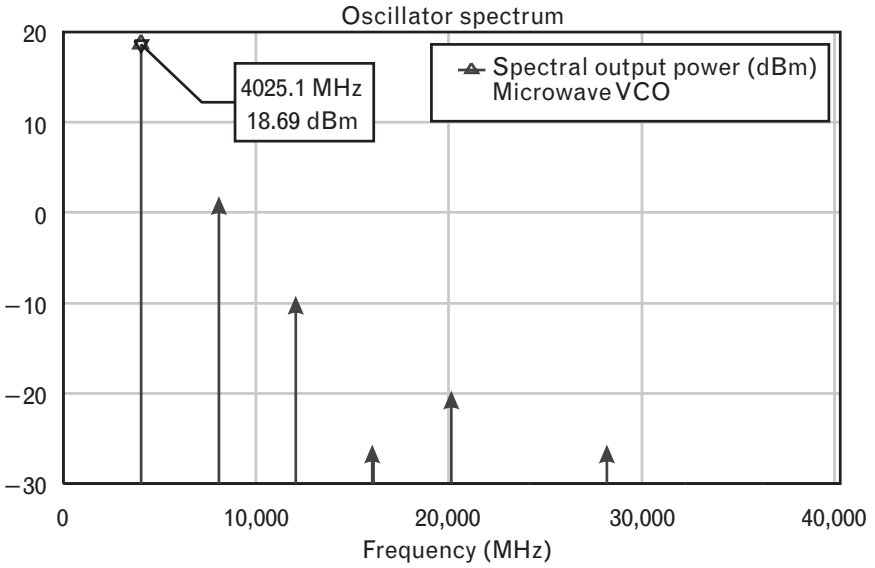


FIGURE 6.72
Oscillator output spectrum with varactor voltage of 2.95V.

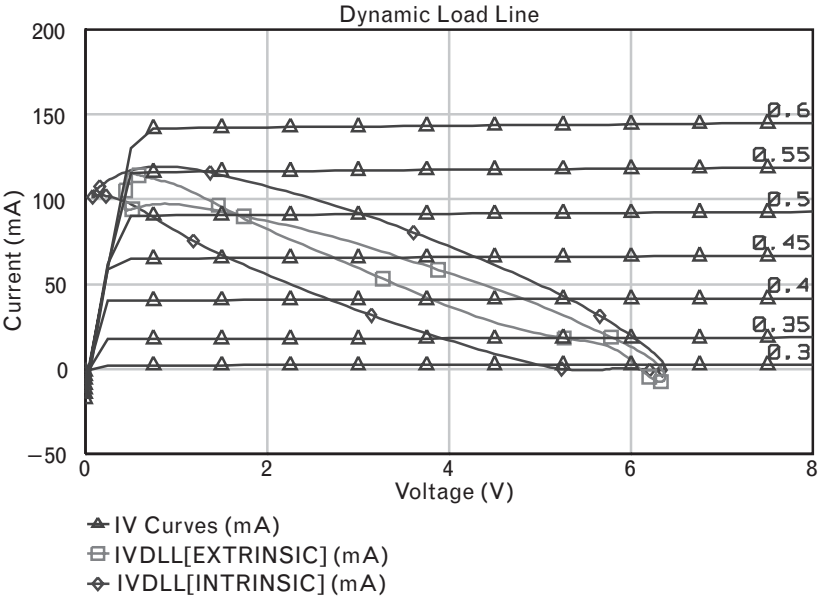


FIGURE 6.73
The load line of the VCO of Figure 6.71.

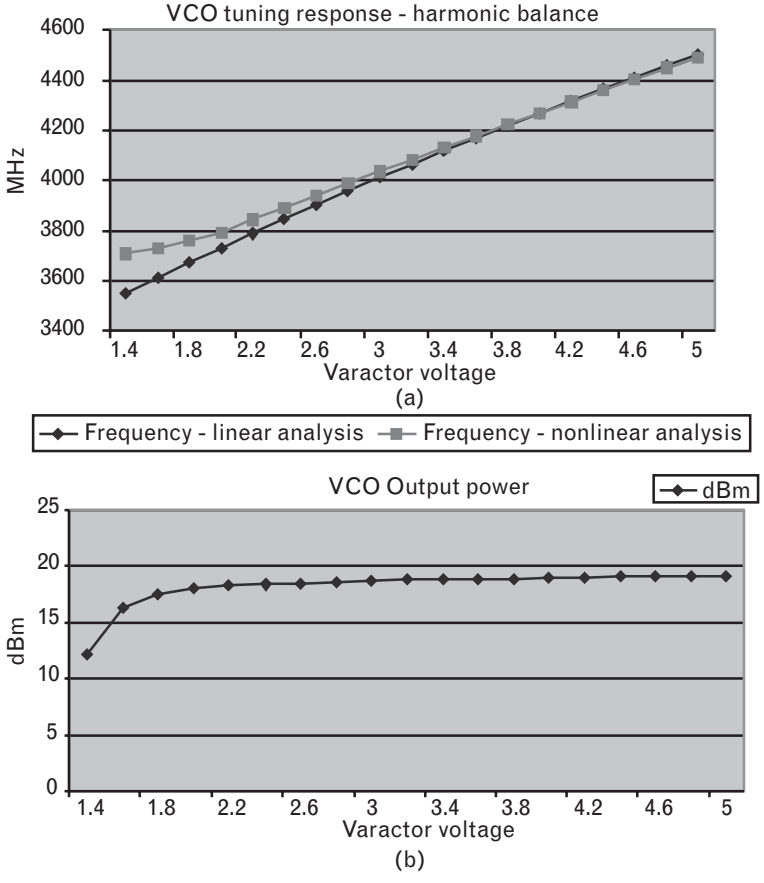


FIGURE 6.74
(a) The tuning curve of the VCO derived from nonlinear simulations, superimposed on the same curve derived from large-signal S-parameter analysis, as in Figure 6.67. (b) The fundamental oscillator output power as a function of varactor voltage.

voltages. It is also quite predictable, since Table 6.4 also indicates that the negative resistance starts to diminish as the frequency is lowered. One way to overcome this problem would be to repeat the design using a nominal design frequency below the 3,600 MHz that we used, to ensure sufficient loop gain and negative resistance exist at lower frequencies, if desired.

6.3 Problems

1. Reconstruct the oscillator of Section 6.1.1.2 using similar *S*-parameters. Instead of optimizing the open-loop system for $G = 1$, optimize for $s_{21} = 1$ instead, with good input and output match. Close the loop and look into the oscillator output port. What is the impedance looking in? Will the device oscillate into a 50-**Ω** load? Why not?
2. Reconstruct the oscillator of Section 6.1.1.2 using similar *S*-parameters.
   (a) This time, replace the resonator with a (lowpass) series L–shunt C–series L section and optimize the open-loop system for $G = 1$ with large-signal *S*-parameters. What is the negative impedance looking into the oscillator output? Is this expected? Now plot the polar Nyquist plot, with both small- and large-signal *S*-parameters. Explain the behavior of the small-signal Nyquist plot by comparing it with the Bode plot. Will this device start to oscillate?
   (b) Replace the resonator with a (highpass) shunt L–series C–shunt L section and repeat the Nyquist plot? Will this device start to oscillate? What are the differences with the circuit in (a)?
3. One oscillator has a Q of 5, another a Q of 50. Which oscillator reaches steady-state conditions first? Which oscillator can be quenched more quickly? Use (6.12) to explain. Are these results intuitive? Can you think of a mechanical system that behaves the same way?
4. Using the small-signal circuit for a transistor, calculate the base impedance of Figure 6.16 as a function of g_m. Show that for certain values of terminating capacitance that it can be negative. Calculate $Z(A)$, and plot its behavior on the Smith chart as g_m is reduced.
5. For the crystal in Section 6.1.3.4, derive the capacitance ratio r and figure of merit M from the equations in Volume I, Chapter 8. What is the calculated frequency shift above the series motional resonance due to the package capacitance? What is the antiresonance frequency?
6. Replot Figure 6.25 on a Smith chart, rather than on Cartesian *R*-*X* axes. What do you notice about the crossing angles? What might

be one advantage of using a Smith chart for the plot rather than Cartesian axes? (*Hint:* Consider the dual circuit, which would require *G*-*B* axes.)

7. (a) Suppose the circuit of Figure 6.31, with a turns ratio of 1:1, has a device resistance of -2Ω, a reactance slope versus frequency of 200Ω/GHz at the resonant frequency of 2 GHz, and resonator losses of 0.4Ω. The load resistor is therefore 1.6Ω. Calculate Q_L, Q_{EXT}, and Q_O.
   (b) Now increase the coupling of the resonator by changing the turns ratio to 2:1 (step up). What must the load resistor be changed to? What are the new values of Q_L, Q_{EXT}, and Q_O? Which circuit has the lowest noise, the highest output power, and the lowest frequency pulling?
   (c) What happens if the coupling is further increased by increasing the turns ratio to 3:1?
8. Derive the open-loop gain expression (6.41) for the Colpitts topology. What are the conditions for startup of oscillation? At steady state, what must the load impedance equal?
9. At 1 GHz, a device has a measured small-signal output impedance of $-15 - j20\Omega$. As the output power increases, the impedance changes to $-10 - j20\Omega$. Plot the *R*-*X* plot. Draw the load line for the circuit that produces large-signal steady-state oscillations with minimum phase noise. Show one circuit diagram that could produce such a load line $Z_L(f)$. What is the Q of your circuit? If the Q of the circuit you have produced is doubled, what happens to the frequency variation with tuning?
10. At 10 GHz, the noise floor of an open-loop system is –170 dBc/Hz, and the corner frequency for the device is 10 kHz. If the system specification on oscillator phase noise is that at a channel spacing of 1 MHz the noise must be better than –140 dBc/Hz, what is the loaded Q of the required oscillator? (*Hint:* Construct the noise plot to determine the 3-dB frequency of the resonator.)
11. Prove (6.65) for the timing jitter of an oscillator.
12. Derive the multiplicative factors 0.891, 0.794, and 0.708 for $|s_{21}|$ to model 1-, 2-, or 3-dB compression of the transistor.
13. The VCO in Section 6.2.2. used a varactor with a zero-bias capacitance of 1.5 pF. We also added in extra inductance at the base. Simulate the achievable tuning bandwidth using the same varactor voltage variation when: (a) $C_j(0) = 0.5$ pF; and (b) $C_j(0) = 4.5$ pF. Both values require the input inductance to be varied. Assume that we still wish the frequency to be 3,700 MHz when the varactor voltage is 1.9V.

References

[1] von Barkhausen, H., *Lehrbuch der Elecktronen-Roehren, Band 3, Rueckkopplung*, Verlag S. Hirzel, 1935.

[2] Clarke, K., and D. Hess, *Communication Circuits: Analysis and Design*, Reading, MA: Addison-Wesley, 1978.

[3] Randall, M., and T. Hock, "General Oscillator Characterization Using Linear Open-Loop S-Parameters," *IEEE Trans. on Microwave Theory and Techniques*, Vol. MTT-49, No. 6, June 2001, pp. 1094–1100.

[4] Meskoob, B., and S. Prasad, "Loop-Gain Measurement and Feedback Oscillator Design," *IEEE Microwave and Guided Wave Letters*, Vol. 2, No. 9, September 1992.

[5] Parzen, B., and A. Ballato, *Design of Crystal and Other Harmonic Oscillators*, New York: Wiley Interscience, 1983.

[6] Nguyen, N. M., and R. G Meyer, "Start-Up and Frequency Stability in High-Frequency Oscillators," *IEEE Journal of Solid-State Circuits*, Vol. JSSC-27, No. 5, May 1992, pp. 810–819.

[7] Kurokawa, K., "Some Basic Characteristics of Broadband Negative Resistance Oscillator Circuits," *Bell System Technical Journal*, Vol. 48, July–August 1969, pp. 1937–1955.

[8] Leenaerts, D., J. van der Tang, and C. Vaucher, *Circuit Design for RF Transceivers*, Boston, MA: Kluwer Academic Publishers, 2001.

[9] Vendelin, G. D., A. M. Pavio, and U. L. Rohde, *Microwave Circuit Design Using Linear and Nonlinear Techniques*, New York: John Wiley & Sons, 1990.

[10] Gilmore, R. J., and F. J. Rosenbaum, "An Analytic Approach to Optimum Oscillator Design Using *S*-Parameters," *IEEE Trans. on Microwave Theory and Techniques*, Vol. MTT-31, August 1983.

[11] Rohde, U. L., "All About Phase Noise in Oscillators," *QEX: ARRL Experimenter's Exchange*, February 1994.

[12] Rohde, U. L., "Designing Low Phase Noise Oscillators," *QEX: ARRL Experimenter's Exchange*, October 1994.

[13] Rohde, U. L., and T. T. N Bucher, *Communications Receivers, Principles, and Design*, New York: McGraw-Hill, 1988.

[14] Leeson, D. B., "A Simple Model of Feedback Oscillator Noise Spectrum," *Proceedings of the IEEE*, February 1966, pp. 329–330.

[15] Verdier, J., et al., "Analysis of Noise Up-Conversion in Microwave Field-Effect Transistor Oscillators," *IEEE Transactions on Microwave Theory and Techniques*, Vol. MTT-44, No. 8, August 1996, pp. 1478–1481.

[16] Llopis, O., et al., "Ultra Low Phase Noise Sapphire—SiGe HBT Oscillator," *IEEE Microwave and Wireless Components Letters*, Vol. 12, No. 5, May 2002, pp. 157–159.

[17] Tanski, W. J., "Development of a Low Noise L-Band Dielectric Resonator Oscillator," *Proc. IEEE Int. Freq. Control Symp.*, pp. 472–477.

CHAPTER 7

Mixers and frequency multipliers

In Volume I, Chapter 3, we explore the importance of mixers in systems and see that the upconversion and downconversion of signals is crucial to radio operation. In some sense, mixers are highly nonlinear components, in which the higher-order terms in a system transfer characteristic are intentionally used to translate between one frequency and another. In another sense, however, the relationship between the input signal and its translated counterpart needs to be quite linear, in which all the usual linear concepts of superposition and matrix algebra apply.

7.1 Mixer overview and their applications in systems

As shown in Figure 7.1, a mixer is a three-port device, which in addition to the input (RF) signal port and output (IF) signal port, uses a third *local oscillator* (LO) port to drive the mixer. This driving action, sometimes called switching or modulation because of its impact on the mixer device(s), is highly nonlinear and causes either the device conductance or transconductance to switch between two states, one with a low (trans)conductance and the other with a high (trans)conductance. We will see in this chapter that almost all mixers use either a time-variant conductance or a time-variant transconductance nonlinearity to achieve frequency translation. We will use the variable $g(t)$ to represent this time-variant nonlinearity.

The switching between the two states occurs at the local oscillator frequency f_{LO}, so the (trans)conductance waveform will contain at least a fundamental component, and possibly higher harmonics as well. If the local

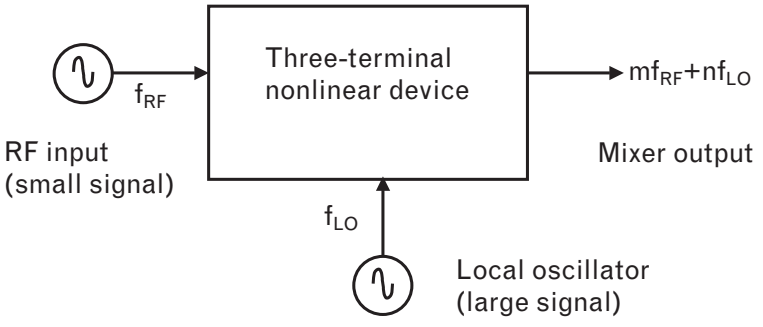


FIGURE 7.1
A generalized mixer model.

oscillator signal is strong enough to cause the device to become nonlinear, then we may write

$$g(t) = g_0 + g_1 \cos(\omega_{LO} t) + g_2 \cos(2\omega_{LO} t) + g_3 (3\omega_{LO} t) + \ldots \quad (7.1)$$

where $\omega_{LO} = 2\pi f_{LO}$.[1] For instance, if the device is switched between an on and a perfect off state, the (trans)conductance waveform is square, with minimum value zero and maximum value corresponding to the on-conductance g_{ON}. In that case, (7.1) would become

$$g(t) = \frac{g_{ON}}{2} + \frac{2g_{ON}}{\pi}\cos(\omega_{LO} t) - \frac{2g_{ON}}{3\pi}\cos(3\omega_{LO} t) + \frac{2g_{ON}}{5\pi}\cos(5\omega_{LO} t)\ldots \quad (7.2)$$

The secret to many mixers, and in fact much of the active research in mixers, is in the baluns or combiners that simultaneously impress the strong LO switching waveform across the mixer added to the much smaller RF input signal, $v_{RF}\cos(\omega_{RF} t)$, where $\omega_{RF} = 2\pi f_{RF}$. In this chapter, the term *balun* is generally used for the three- or four-port device that is configured to linearly sum the incident voltages at the two balun input ports (the LO and RF), rather than for achieving single-ended to differential conversion as is commonly the case in other types of circuits (e.g., push-pull amplifiers). Of course, the same circuit can often be used for either function. The reader is referred to [1, 2] for excellent material on baluns.

The effective voltage applied across the time-varying conductance is then the input signal voltage. For although the mixer model in Figure 7.1 shows three ports, diodes have only two terminals and transistors three, so some means of feeding the device with two signals and for extracting the third needs to be created. If this can be done, then the output current of interest is simply

$$\begin{aligned} i(t) &= g(t)v(t) \\ &= (g_0 + g_1 \cos(\omega_{LO} t) + g_2 \cos(2\omega_{LO} t) + g_3 \cos(3\omega_{LO} t) + \ldots) v_{RF} \cos(\omega_{RF} t) \\ &= g_0 v_{RF} \cos(\omega_{RF} t) + \frac{g_1}{2} v_{RF} \left[\cos(\omega_{LO} - \omega_{RF})t + \cos(\omega_{LO} + \omega_{RF})t\right] \\ &+ \frac{g_2}{2} v_{RF} \left[\cos(2\omega_{LO} - \omega_{RF})t + \cos(2\omega_{LO} + \omega_{RF})t\right] \\ &+ \frac{g_3}{2} v_{RF} \left[\cos(3\omega_{LO} - \omega_{RF})t + \cos(3\omega_{LO} + \omega_{RF})t\right] \end{aligned} \quad (7.3)$$

1. In this chapter we shall assume the reader is equally comfortable with either ω or f to express frequency, since the former is simpler to use in trigonometric expressions such as *sine* and *cosine*. We shall alternate freely between using both.

The RF signal has been translated in frequency, and its phase and amplitude are preserved in the Fourier components of the output current waveform. In theory, the LO carrier has been suppressed at the output. However, the expression implies that we must carefully consider the harmonic embedding impedances of the diode in order to preserve the conductance relationship and to select the desired output components. For a downconverter, we are generally interested in the IF component at radian frequency $\omega_{LO} - \omega_{RF}$, the difference between the local oscillator and the RF component. For an upconverter, it is the IF component at frequency $\omega_{LO} + \omega_{RF}$ that is of interest. The amplitude of the desired IF current component is then $(g_1/2)v_{RF}$, which is linearly related to the input RF signal strength.

As a rule, we should consider at the very least the IF, RF, and LO embedding (or matching) impedances of the device. Often, the higher harmonic current components will be short-circuited by the parasitic impedances of the device itself, although not always. As long as (7.1) is not corrupted by doing so, short-circuit embedding impedances at unwanted frequencies are generally preferred because they will prevent any unwanted distortion voltages that could arise from remixing within the device, and result in better intermodulation performance. We discuss the impact of these terminations later on, in Section 7.2.5.2.

Recall that in (7.1), the incremental conductance g is defined as $\partial I/\partial V$ or $\partial I_O/\partial V_{IN}$ in the case of transconductance. Now in the simple case of a square-law device, the total input current to the device I is expressed as a second-order power series of the total voltage V across it,

$$I = I_Q + G_1 V + G_2 V^2 \tag{7.4}$$

so that

$$g(t) = G_1 + 2G_2 V(t) \tag{7.5}$$

If we let $V(t) = V_{LO}\cos(\omega_{LO}t)$, then comparing (7.1) and (7.5) gives in this case $g_0 = G_1$ and $g_1 = 2G_2 V_{LO}$. Thus, in this simplest case:

- The desired IF component $(g_1/2)v_{RF}$ depends on the second-order nonlinearity G_2 in the transfer characteristic of the device. This makes modeling of mixers more difficult than amplifiers, where the fundamental output depends instead principally on G_1, the linear (trans)conductance term.
- The desired IF component of current is linearly related to the input signal v_{RF} (in both amplitude and phase).

In higher-order devices, differentiation of the equivalent of (7.4) produces terms in $(n+1)G_{n+1}V(t)^n$, which upon expansion of the LO voltage term $V^n_{LO}\cos^n(\omega_{LO}t)$ into its harmonic components will produce additional components $V^k_{LO}\cos(k\omega_{LO}t)$, $k = 0,1\ldots n$ that will change the values of g_0, g_1, and so on, and introduce additional dependency on the LO signal level. However, the principle of (7.3) still stands, so that sum and difference frequencies will flow in the mixer current and the difference frequency component will still be linearly related to the input RF voltage.

We have overlooked one assumption that is not quite negligible, and that is that the RF voltage itself is part of the total applied voltage in (7.4). Although generally negligible compared to the much larger LO voltage, it will, in fact, be impressed across the device and as a result, $g(t)$ in (7.5), and the coefficients g_0, g_1, and so on will also have a weak dependency on the RF signal. This introduces harmonic terms in the RF frequency ω_{RF} in a similar way to ω_{LO}, as well as introduces dependency on the magnitude and phase of the RF voltage, so that the mixer now shows nonlinear dependence on the RF component. Thus, in general, the output current of a mixer will contain terms at frequencies

$$m\omega_{RF} \pm n\omega_{LO} \quad m, n = 0,1\ldots \tag{7.6}$$

Provided the LO voltage is much stronger than the RF voltage, the output current term at the difference frequency is linearly related in amplitude and phase to the input RF signal. This frequency, the IF component, has $m = n = 1$.

As shown in Figure 7.2, the LO can be either below the RF band of interest, in which case the mixer is referred to as a low-side downconverter, or above it, resulting in a high-side downconverter. The difference between the IF in a high-side downconverter and the IF in a low-side downconverter is that the phase of the two IF signals will be 180° apart. In the second case $\omega_{LO} - \omega_{RF}$ will be positive and in the first case will be negative, since if $\omega_{LO} > \omega_{RF}$, then $\sin(\omega_{RF} - \omega_{LO})t = \sin[(\omega_{LO} - \omega_{RF})t + \pi]$.

The implications of the "±" term in (7.6) are important. In downconverters, it implies that any undesired RF components at an image frequency of $\omega_{LO} - \omega_{IF}$ (for LO below the desired RF) or $\omega_{LO} + \omega_{IF}$ (for LO above the desired RF) will also be downconverted to the IF. Figure 7.2 illustrates this case. We saw in Volume I, Chapter 3, that the downconversion of an image frequency has implications for both the system noise floor and spurious response.

In upconverters, the "±" term implies that the signal is mixed to both a lower and an upper sideband, as shown in Figure 7.3 for upconversion of the IF to RF. Note that an upconverter can be either a sum or a difference mixer, depending on the sideband selected. Although upconverters are

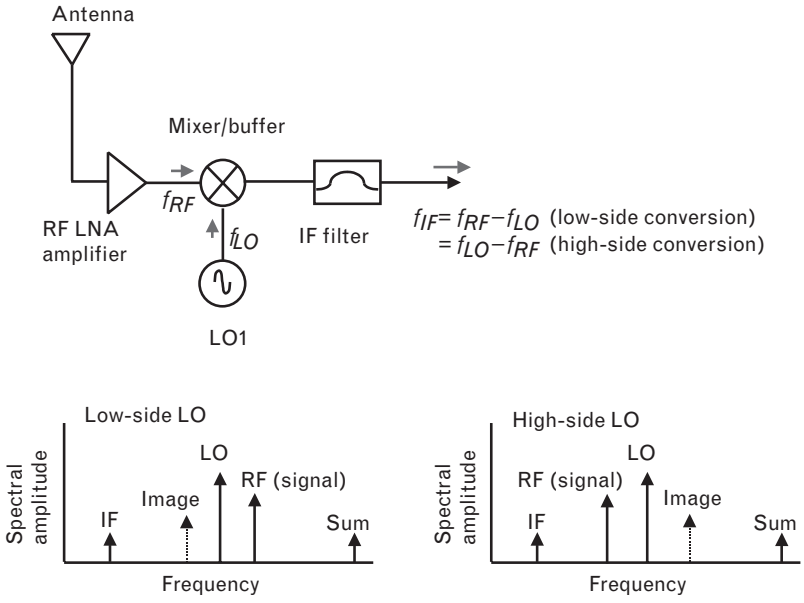


FIGURE 7.2
A mixer downconverter system.

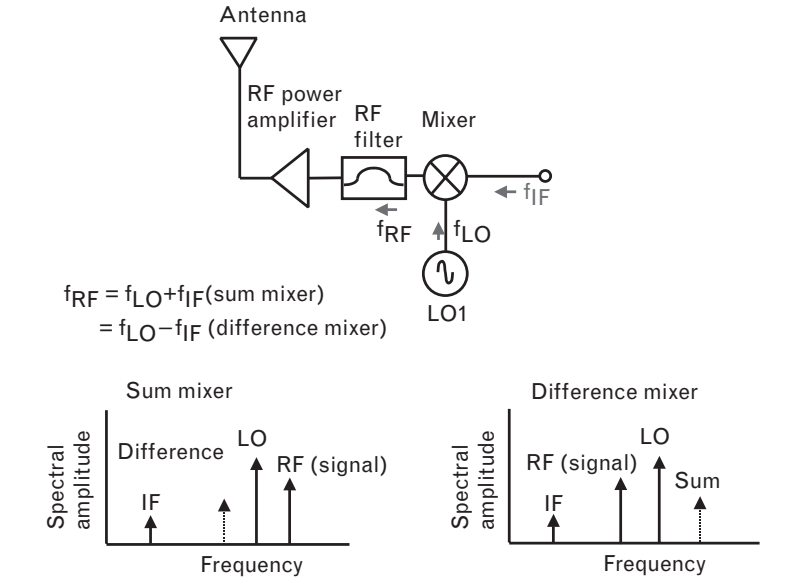


FIGURE 7.3
A mixer upconverter system.

frequently referred to as modulators in a transmitter (as shown in the figure), upconverters are also useful in receivers where the RF band covers a large percentage bandwidth and downconversion would require a large percentage tuning range for the VCO, or cause problems with the image frequency lying in band.

As we see in Volume I, Chapter 3, mixers are characterized by comparing the relation between the output current, generally at the IF

frequency,[2] to the input RF voltage. In terms of power, the conversion gain is defined for a mixer as simply

$$\text{Conversion gain} = \frac{\text{Output signal level (IF)}}{\text{Input signal level (RF)}} \tag{7.7}$$

For passive mixers, such as diode mixers, there is always conversion loss. For instance, (7.2) and (7.3) result in a term in the IF frequency current of $g_1 v_{RF}/2 = g_{ON} v_{RF}/\pi$, while the RF frequency current is $g_0 v_{RF} = g_{ON} v_{RF}/2$. The ratio of the respective currents is therefore $2/\pi$ so that the conversion gain is $(2/\pi)^2 = 0.41 = -3.92$ dB if the impedances are equal. Because the gain is negative in a passive mixer, we commonly refer to the conversion loss L_C instead, the inverse of (7.7).

The minimum theoretical conversion loss in any passive mixer is 3.92 dB, in which the device is switched with a square wave (i.e., one with a large local-oscillator signal that saturates the device). The loss is invariant to the number of devices in the mixer, since the IF and RF currents will always flow in each device with the same ratio. Of course, any mismatch at the RF port or the IF port will make the conversion loss worse, since the square of the ratio of currents only equals the power ratio when the impedances are equal.

In an ideal mixer, we see from (7.2) and (7.3) that the amplitude of the higher harmonic responses of the LO simply falls as the Fourier coefficients of a square wave. The gain of the IF is –3.92 dB, that of $2\omega_{LO} - \omega_{RF}$ is $(2/3\pi)^2 = -13.5$ dB and that of $3\omega_{LO} - \omega_{RF}$ is $(2/5\pi)^2 = -17.9$ dB. These are the potential spurious responses of the mixer that we study in Volume I, Chapter 3.

In general, the conversion loss will become worse as the LO signal weakens. We can see this from (7.2), because if the LO signal is insufficient to drive the conductance as a square wave, but instead drives it *sinusoidally* between the same two peak states, then the $g_1 v_{RF}/2$ term for the IF frequency component in the output current in (7.3) becomes $g_{ON} v_{RF}/4$ (rather than $g_{ON} v_{RF}/\pi$). The power ratio in (7.7) then becomes 0.25 or –6 dB. As the LO becomes even weaker and is unable to drive the conductance between an off-state and a fully saturated on-state, the peak value of the IF current becomes correspondingly smaller and the conversion loss and noise figure become worse.

It is sometimes convenient to model a mixer as a switch. A number of output waveforms are shown in Figure 7.4 for various switching configurations. In the ideal multiplier of Figure 7.4(a), the sinusoidal LO signal of frequency 1 GHz and the (equal level) RF signal at 1.1 GHz multiply to

2. IF, of course, means intermediate frequency, so "IF frequency" is syntactically redundant. However, like many other texts, we prefer the redundancy of this term or of "RF frequency" to distinguish it from "IF" or "RF," which on its own can typically refer to the signal component (i.e., voltage, current, or power).

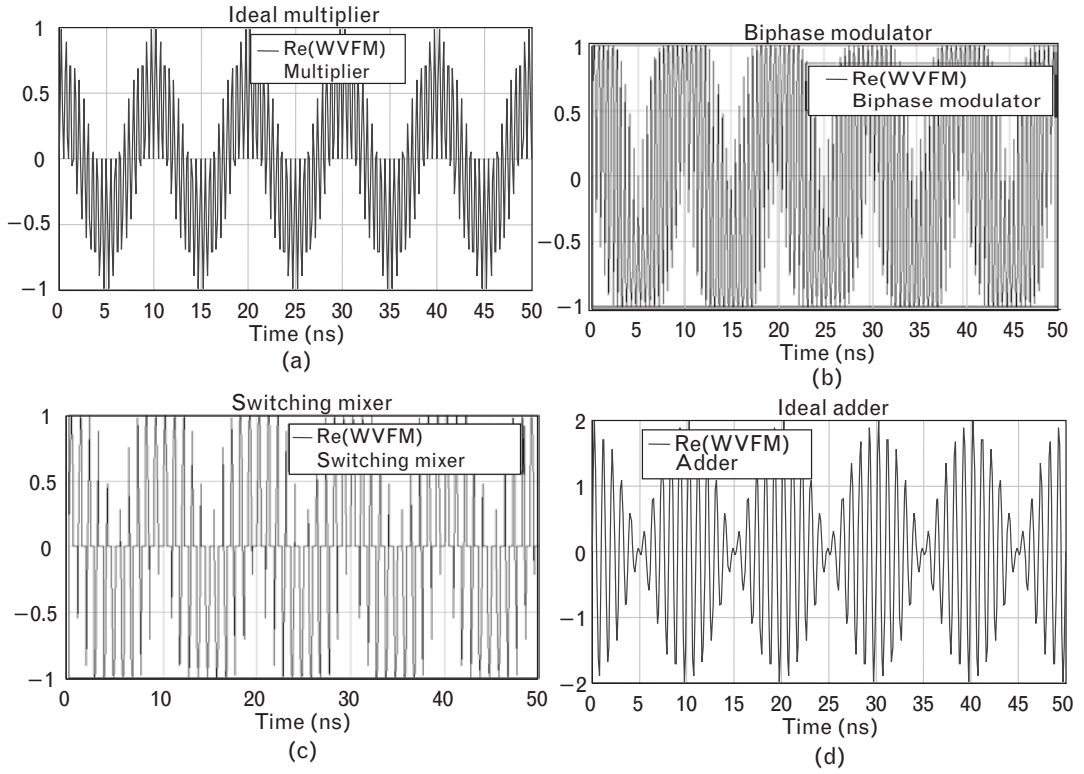


FIGURE 7.4 *Waveforms for various outputs from combining equal level 1-GHz RF and 1.1-GHz LO signals, from (a) a linear (ideal) multiplier of the two sinusoids, (b) a biphase modulator with phase switched at the LO rate, (c) a switch opened and closed at the LO rate, and (d) an ideal adder of the two sinusoids.*

produce a 100-MHz IF output. In the biphase modulator of Figure 7.4(b), the LO signal is a square wave that multiplies the RF component by either +1 or −1 (when the RF phase is inverted). In the switch of Figure 7.4(c), the LO multiplies the signal by either +1 (closed switch) or 0 (open switch). In all cases, the IF component with period 10 ns is clearly visible. As we just calculated, the IF voltage in Figure 7.4(b) with the square-wave switching LO waveform is larger by a factor of $4/\pi$ compared with the sinusoidal LO of equal peak amplitude in Figure 7.4(a). For comparison, the result of a simple linear addition of the two tones is shown in Figure 7.4(d). Here, the modulation envelope is one-half the IF frequency, since

$$\cos(\omega_1 t) + \cos(\omega_2 t) = 2\cos\left[\frac{\omega_1 t + \omega_2 t}{2}\right]\cos\left[\frac{\omega_1 t - \omega_2 t}{2}\right] \qquad (7.8)$$

Finally, we need to keep in mind that the discussion of conversion gain is with reference to a single-sideband system, for which we translate only

one RF component to the IF. Many digital radio systems transmit only the upper or lower sideband of an upconverted signal to preserve spectrum. However, some systems such as analog AM or FM radios use both. In such a double sideband system, then, there will, in fact, be two RF signals that are downconverted to the IF frequency in the receiver, one at $\omega_{RF} = \omega_{LO} - \omega_{IF}$ and the other at $\omega_{RF} = \omega_{LO} + \omega_{IF}$. In that case, the conversion gain and the IF component in (7.3) will be double compared with a single sideband system, where there is only one RF component and a second *null* sideband, then known as the image frequency. Similarly, the double-sideband noise figure is up to 3 dB improved (smaller) compared with the single-sideband noise figure, since the IF noise is similar for both mixers but the signal is twice as large for *double sideband* (DSB) operation.

In addition to the degradation in system noise figure introduced by the conversion loss L_C of a mixer, noise sources within the mixer device itself further corrupt the noise figure. For instance, the effect of $1/f$ noise in MESFETs can be severe if the IF frequency is below the corner frequency of the flicker noise (normally less than 1 MHz), as this noise will add to the output. If t_r is the ratio of the measured noise power at the IF output compared with the input thermal noise in the same measurement bandwidth, then the mixer noise figure is given by

$$F = t_r L_C \tag{7.9}$$

In Volume I, Chapter 3, we study the expression for cascaded noise figure and find that if a mixer is preceded by a high gain, low-noise amplifier, then the cascaded noise figure of the system is set principally by the amplifier itself. For a two-component system, the cascaded noise figure is simply

$$F = F_1 + \frac{F_2 - 1}{G_1} \tag{7.10}$$

so that if G_1 is sufficiently high (as in an LNA) the second term (from the mixer) can be neglected. Some caution is needed here, however, because we saw that in the case of a receiver, it can be good practice to insert a second RF filter following the amplifier and prior to the mixer. This is because if the amplifier is reasonably broadband (as many LNAs are), then the gain at the image frequency will be similar to the gain at the RF signal frequency. Therefore, in a broadband mixer, the noise floor at the image frequency will fold onto the RF signal noise floor when downconverted to the IF, resulting in a 3-dB loss in system sensitivity, no matter how good the preceding component noise figure. The purpose of the preceding RF filter should therefore be to remove as far as possible the effect of the image noise. Alternatively, an image-reject mixer can be used to automatically

reject the image component from the IF output. Such a mixer is an important component in single-sideband operation where the desired RF is present on only one side of the local oscillator and an unknown image on the other. We will discuss such a mixer later in this chapter.

If we consider the case of the mixer followed by a narrowband IF amplifier that selects and amplifies the desired IF component, as in Figure 7.5, then (7.10) becomes

$$\begin{aligned} F &= t_r L_C + \frac{F_{IF} - 1}{1/L_C} \\ &= L_C\left(t_r + F_{IF} - 1\right) \end{aligned} \tag{7.11}$$

If the noise contribution of the mixer device itself can be neglected, then $t_r = 1$ and $F = L_C F_{IF}$, that is, the cascaded noise figure in decibels is just the conversion loss of the mixer plus the noise figure of the IF amplifier.

It is also common to refer to the *linearity* of a mixer. Although a highly nonlinear device, the relationship between the relatively low-level RF signal and the IF signal is linear over some range. In this respect, terminology is borrowed from an amplifier, and the linear gain (conversion loss), 1-dB compression point, and intercept point can all be defined. However, the output power of interest is the IF power, and the input power against which the linearity is measured is the RF input power. The LO power level at which the measurements are made is usually fixed and needs to be sufficiently high to achieve the desired conductance waveform. The mixer equivalent of the harmonic components in an amplifier are the (m,n) distortion products that occur at frequencies as given in (7.6) (i.e., by $m\omega_{RF} \pm n\omega_{LO}$ $m,n = 1,2,\ldots$) The *third-order intercept point* (IP3) in a mixer is defined by the extrapolated intersection of the primary IF response with the two-tone third-order intermodulation IF product that results when *two* RF-signals are applied to the RF port of the mixer. These are at frequencies of $(2\omega_{RF1} - \omega_{RF2}) - \omega_{LO}$, and $(2\omega_{RF2} - \omega_{RF1}) - \omega_{LO}$ for a low-side downconverter.

As for an amplifier, it is common to model the output IP3 point as 10 dB above the output 1-dB compression point. This result, which was given for a general third-order nonlinearity in Volume I, Chapter 3, has equal

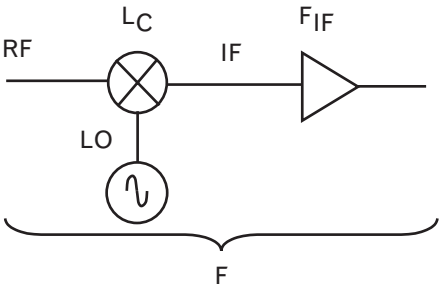


FIGURE 7.5
A mixer followed by an IF amplifier, for noise calculation.

applicability here because the intermodulation distortion still results from mixing within the cubic term in the (trans)conductance nonlinearity. As with amplifiers, it is only a rule of thumb, and in real devices, higher intercept points can be achieved if higher-order terms in the nonlinearity offset the third-order term. We will see examples of mixers with better than 10-dB differences later in this chapter. Figure 7.6 shows these definitions graphically and the mixer *spurious-free dynamic range* (SFDR) corresponding to those input powers for which the output power is above the noise floor but still free from third-order distortion products.

In the following sections we will look at various mixer implementations and the different technologies that are used to design them.

7.2 Diode mixers and their topologies

Because of their simplicity, broadband coverage, and lack of need for dc power, diode mixers are ubiquitous. They may be used singly, combined in pairs, quads, and even octets, with an array of different baluns to support either single-ended or differential inputs and a variety of different power levels. Even though diodes have been replaced by newer devices in many other solid-state applications, diode mixers remain fundamental to frequency conversion.

The three basic topologies of diode mixers are shown in Figure 7.7. The single-ended mixer uses a single diode, and the bulk of the work lies in designing the filters to ensure adequate decoupling between the RF, LO, and IF ports, and to ensure the RF voltage is impressed across the diode and that the IF current can be extracted. The single-balanced mixer uses a

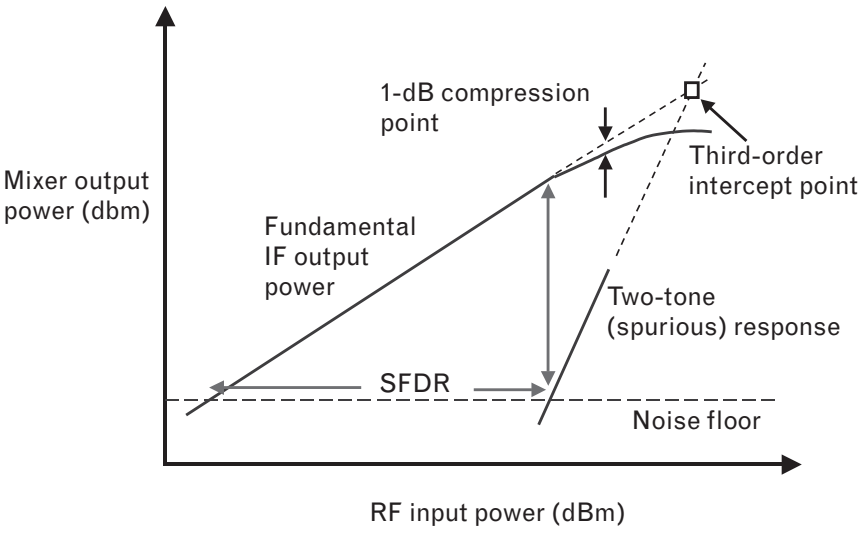


FIGURE 7.6
The third-order intercept point, 1-dB compression point, and spurious-free dynamic range (SFDR) for a mixer.

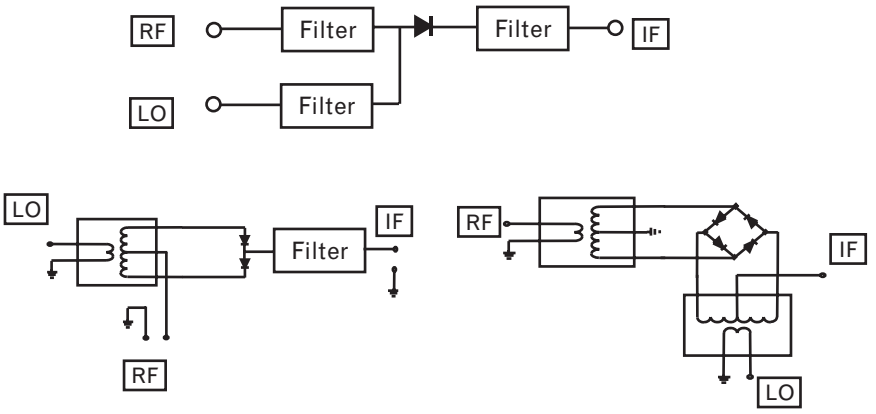


FIGURE 7.7
Single-ended, single-balanced, and double-balanced diode mixer topologies. (After: [1].)

balun to achieve these constraints and to build in RF-LO isolation, while the double-balanced mixer supports differential RF and LO inputs and ensures even better isolation between the RF and LO.

7.2.1 Single-ended mixer

Figure 7.8 shows a single-ended mixer in a downconverter configuration. The input balun, through which the RF and LO voltages are impressed across the diode, is shown schematically as a directional coupler, in which both the incident RF and LO voltage waves are coupled to the output port of the coupler, and the reflected wave from the mixer is directed to the fourth, terminated port. The principle of the mixer is that the LO voltage, impressed across the anode of the diode, is large enough to switch the diode on and off, according to (7.1). The RF signal, which in this representation needs to be close to the LO in frequency since it shares a common input coupler, should be matched to the input of the diode so that the RF voltage $v_{RF}\cos(\omega_{RF}t)$ is impressed across the diode and the output current can assume the form (7.3). The RF choke schematically represents the input matching network; and the lowpass filter at the output effectively short-circuits the cathode to ground at the RF and LO frequencies, so that both input signals are fully impressed across the diode itself.

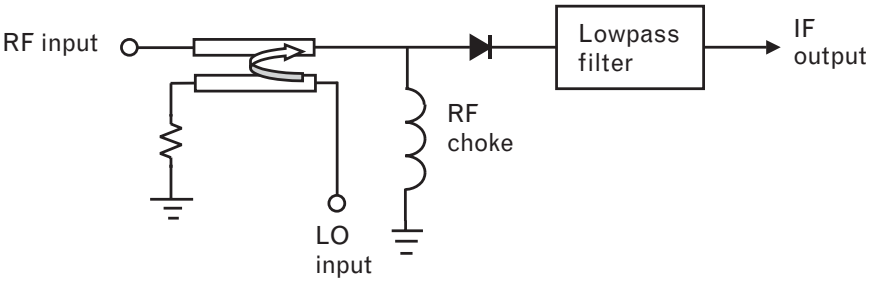


FIGURE 7.8
Basic topology for a single-ended mixer.

The concept of an impedance match at a particular frequency requires some thought. For instance, (7.2) shows that the impedance of the diode is a time-varying quantity, and nonlinear as well. Nonetheless, the impedance at a particular frequency (say, the fundamental LO) is simply the coefficient of the term at the relevant (LO) frequency. This is a single constant quantity at any particular drive level, although still nonlinear in that it will change with LO drive level. A similar constant relationship between the voltage and current components of the time-varying conductance at the RF or IF frequencies will also exist, yielding the equivalent impedances at these frequencies for matching.

The IF current is extracted from the output (cathode) of the diode through a lowpass filter, which also eliminates the unwanted, high-frequency components. The input (anode) of the diode needs to be short-circuited at low frequencies so the IF voltage is developed across the diode itself. The RF choke achieves this, and in addition provides a low resistance path for the rectified LO current that will flow. If the RF choke were not present, a negative dc voltage would develop across the anode of the diode and the nonlinearity of the switching action would be impeded.

The LO drive level can be made quite large for such a mixer, and the usable frequency and bandwidth is determined by the external filters, since the diode itself can be made arbitrarily small to operate up to the millimeter-wave frequencies. Frequently, the output third-order intercept point is approximately equal to the LO power level, and possibly a few decibels higher.

Some configurations of the single-ended mixer use an antiparallel diode pair in place of the single diode in Figure 7.8. This doubles the LO frequency because of the full-wave rectification effect, and somewhat simplifies the filtering requirements for isolation since the fundamental LO frequency can then be set closer to one-half of the RF frequency, using its second-harmonic for mixing. Such a circuit is commonly used at millimeter-wave frequencies. However, the LO drive requirement is about 9 dB higher to achieve the same power at what is now the second harmonic, so if the LO drive is only the same as for a single diode mixer, the input intercept point is approximately 9 dB worse.

FETs can also be configured as diodes and used in their place, for instance in GaAs ICs. The FET has a higher third-order intercept point than the diode in an equivalent circuit, although resistive FET mixers using the variable conductance between drain and source are more common because they are more linear and more reliable.

Single-ended mixers are cheap and simple. They are used in low-cost detectors, for instance in domestic motion detectors. Their greatest drawback is that their ability to prevent radiation of the local-oscillator signal back into the RF port and out of the antenna depends entirely on the selectivity of the input balun. In the case of a microstrip directional coupler, the isolation between the two adjacent ports will rarely be better than 20 dB.

Given that the LO signal is much stronger compared to the RF, a substantial amount of LO radiation can leak out the antenna and advertise the presence of the mixer to all other receivers.

7.2.2 Single-balanced mixer

A single-balanced mixer uses two diodes connected back-to-back as shown in Figure 7.9. The figure shows the RF signal connected to the sum port of a 180° hybrid coupler and the LO signal connected to the difference port, although they can be interchanged. This could correspond to the center tap and the differential input of a coupling transformer, respectively, as shown in Figure 7.9(b). An RF-choke (not shown) is required in shunt at the input to each diode to provide a dc and IF ground if these are not part of the input transformer circuit itself. For instance, in microstrip they would be implemented by short-circuited quarter-wave lines at the RF frequency, to ground the respective input of each diode at the IF frequency and dc. (Note that we are referring to the diode input rather than the anode or cathode now, as each diode should be considered as a mixing element with input and output terminals.) In the case of a downconverter, a short-circuiting lowpass filter is used at the output to ensure that the RF and LO voltages are fully impressed across the diodes themselves. Be careful! A lowpass filter can also attenuate high frequencies by presenting an open-circuit, and this is the wrong impedance level to present to the diodes at the output.

Consider the instantaneous phasing of the LO and RF voltages as shown in the figure. There the anode of D1 is driven instantaneously positive by the LO, and the cathode of D2 is driven negative by the LO. If, in Figure 7.9(c), we represent the LO voltage by a phasor rotating at frequency ω_{LO}, then the LO voltage appears across the two diodes in the (opposing) sense indicated. But because the two diodes are reverse connected, both are turned on and off at the same time, so their conductance

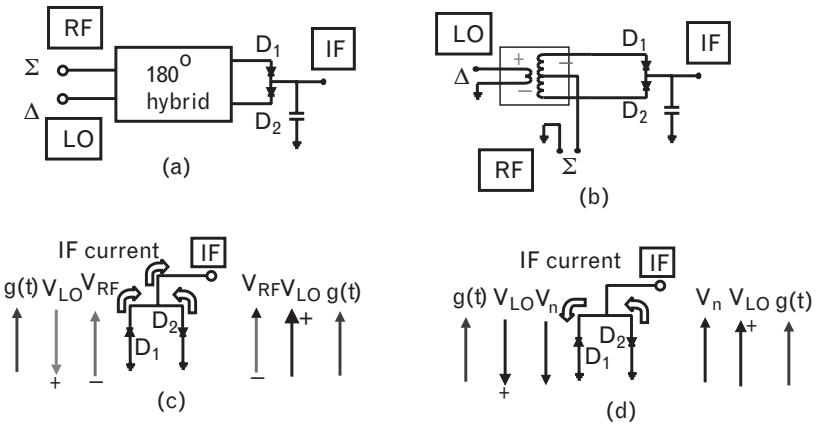


FIGURE 7.9
(a) Schematic for a single-balanced mixer; (b) transformer implementation; (c) phase relationships showing how the IF currents sum at the output; and (d) phase relationships showing AM noise cancellation. (After: [1].)

phasors $g(t)$ are in phase. The RF is fed through the center tap, and if it is instantaneously negative, the anode of D1 and the cathode of D2 are both driven incrementally more negative: The RF phasors will both be in phase and in the sense indicated in Figure 7.9(c). Now, since from (7.3) the IF current is proportional to the product of the conductance and the RF voltage, the IF currents in both diodes flow in the same sense, both exiting the mixer through the IF output port.

Figure 7.9(d) shows the AM noise cancellation property of the single-balanced mixer. If the local oscillator voltage has AM noise V_n, then it occurs simultaneously across the two diodes, in the same sense as the LO itself. Now, the LO noise currents and the conductance phasors have opposite sense in the two diodes, so that when one IF noise current is positive in one diode, it is negative in the other. The noise component of IF current generated from the LO simply circulates between the two diodes: LO AM noise at the output cancels.

A single-balanced mixer can also be constructed using a 90°, or hybrid coupler. The principle here is shown in Figure 7.10, where the phasor representation of the RF and LO voltages are shown. If at some instant of time the RF and LO both have the same phase at their respective input ports, then the RF is delayed 90° at the remotely coupled output port relative to the near coupled port; and conversely, the LO is delayed in the same sense. Thus, in the top leg of the mixer the RF lags the LO by 90°, and in the bottom leg it leads it by 90°. Consequently, the phase difference in the two legs of the mixer is reversed and the differences in the LO and RF phasors are identical to those in Figure 7.9(c). As before, the IF currents flow in phase at the connection between the two diodes.

The hybrid coupler does not bestow on the single-balanced mixer the same advantages as it does the balanced amplifier, because the coupler is driven at two ports with two different frequencies at quite different power levels. Isolation between the two input ports is therefore not as good as in

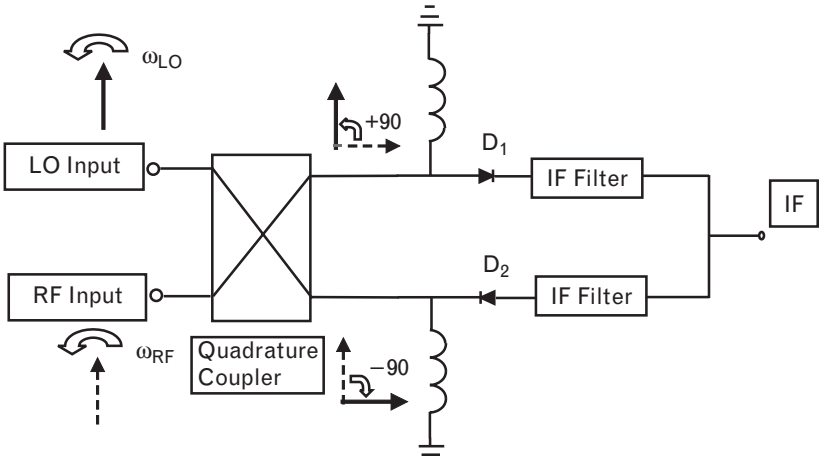


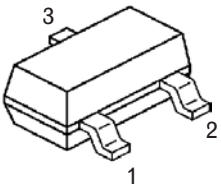
FIGURE 7.10
A single-balanced mixer using a 90° coupler as a balun.

the balanced amplifier, since incident LO power reflected from the two diodes comes out the RF port, and reflected RF power out the LO port. Balance is thus harder to maintain, particularly if the RF or LO driver stages have a poor output match themselves.

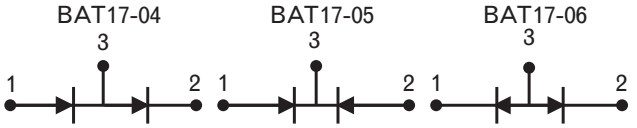
Figure 7.11 shows various combinations of packaged silicon Schottky diodes for such applications. The leads can be configured for direct connection to external baluns for mixer applications (BAT 17-04). Such diode packages are useful for VHF and UHF frequency ranges. Noise figure is about 6 dB, and the diode capacitance is less than 0.75 pF. The typical

FIGURE 7.11
Extract of a data sheet for the Infineon Technologies BAT17 Schottky diode pairs. (Courtesy Infineon Technologies.)

Silicon Schottky diode
·For mixer application in VHF/UHF range
·For high-speed switching application



BAT17



Electrical characteristics at T_A = 25°C, unless otherise specified

Parameter	Symbol	Values			Unit
		min.	typ.	max.	
DC characteristics					
Breakdown voltage $I_{(BR)}$ = 10 μA	$V_{(BR)}$	4	-	-	V
Reverse current	I_R	-	-	0.25	μA
V_R=3		-	-	10	
V_R=4V		-	-	1.25	
V_R=3V, T_A =60°C					
Forward voltage	V_F	200	275	350	mV
I_F=0.1mA		250	340	450	
I_F=1mA		350	425	600	
I_F=10mA					
AC characteristics					
Diode capacitance- V_R=0V, f=1MHz	C_T	0.4	0.55	0.75	pF
Differential forward resistance I_F=5mA, f=10kHz	R_F	-	8	15	Ω

reverse current (at room temperature) is less than 0.03 μA at reverse voltages up to 4V.

The BAT17 forward voltage is 0.34V at 1-mA forward current. This low forward voltage, related to the barrier potential of the metal-semiconductor junction, is typical of well-designed silicon mixer diodes and implies lower LO power is needed to switch the diode than for higher barrier devices. For instance, a high-barrier device requiring 0.7V for the same current would require about 6 dB more LO power than this one. Typical LO power requirements for the BAT17 are between 3 and 7 dBm input power, and as a rule-of-thumb, the *input* 1-dB compression point is about 6 dB below the LO power. The *input* third-order intercept point for diode single-balanced mixers is at best 9.5 dB higher than the LO power, but typically may only be of the same order. As just noted, the intercept point can be improved by using a higher barrier diode instead, made for instance from GaAs. However, silicon diodes have noise corner frequencies around 100 kHz, while those of GaAs are approximately 500 kHz. This could be a problem if the mixer is used at very low IF frequencies, since its $1/f$ noise could fall in-band.

To implement a single-balanced mixer at higher frequencies, individual "beam-lead" diodes can be mounted directly onto microstrip, or chip diodes inserted directly into via holes through the microstrip for connection of one side to ground. Figure 7.12(a) shows two chip diodes connected to a four-port balun known as a microstrip rat race. The diodes are mounted asymmetrically to ground, similar to the transformer topology of Figure 7.9(b). The LO is applied to the difference port at the bottom of the figure, and the RF to the sum port at the left of the figure. The diodes are connected equidistant from the ring by microstrip lines, with added stubs for matching the diode impedance to 50Ω for the RF and LO. By symmetry, it can be seen that the RF is fed to both diodes in phase. The LO is fed out of phase to both diodes. Intuitively, this can be explained because the distance clockwise around the ring from the LO injection point to the first

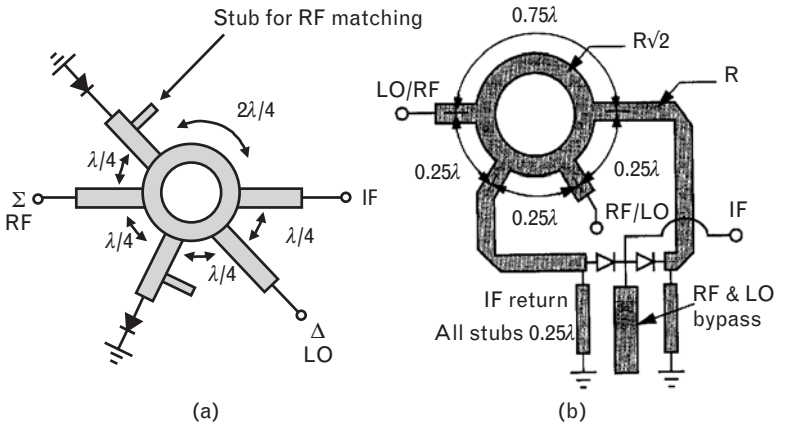


FIGURE 7.12
(a) The rat-race mixer for use at microwave frequencies. (b) An alternate implementation where the IF is extracted directly from the diodes. (From: [2]. © 1998 Artech House, Inc. Reprinted with permission.)

diode arm is one-quarter wavelength, and to the second it is three-quarter wavelengths, a difference of half a wavelength or 180°. (The difference is the same independent of direction around the ring.) The rat-race balun is therefore a 180° balun and the connectivity matches the configuration of Figure 7.9(a). The IF currents sum symmetrically out the port at the right of the figure, with an output filter preventing leakage of RF or LO through this port.

The two diodes can also be mounted back-to-back in the same location (as long as the transmission line lengths from the rat race are the same), as shown in Figure 7.12(b). The IF is then taken directly from the junction of the two diodes, at which an RF/LO short circuit is required, as before, to ground the higher frequencies. The rectified dc current that results will simply circulate in the two diodes, and an IF return at the input to each diode is still required to short-circuit the diode inputs at the IF frequency. Again, the similarity with Figure 7.9(a) is apparent.

It can, perhaps, be seen qualitatively from the reversal of the diodes and the symmetry of the single-balanced structure that the spurious mixing products of $m\omega_{RF} \pm n\omega_{LO}$ in (7.6) will cancel when m and n are both even [i.e., the (2,2), (4,4) products, and so forth]. This can, in fact, be proven using either the phasor representation we used earlier or through a simple mathematical argument as follows. Assume the current in the first diode (from the anode to the cathode) is given by

$$I = I_Q + G_1 V + G_2 V^2 + \ldots \tag{7.12}$$

where V is the anode voltage of the diode and we have omitted higher-order terms for simplicity. This is also the current in the second diode, if the voltage is again measured across the anode and the current convention is into the anode.

Now, if the fundamental signal applied to the input of one diode is $+V$, then the signal applied to the input of the second diode is $-V$ in the case of the LO and $+V$ in the case of the RF. But the input of the second diode is its cathode, thus its LO anode voltage will be $+V$ and RF anode voltage $-V$ [e.g., as in Figure 7.9(c)]. But even-order terms produced in the diodes from the fundamental of the RF or LO will always involve current components in each diode like $(+V)^2$ or $(-V)^2$, which are both positive. Furthermore, the inputs and outputs of the diodes are of opposing polarity, and the direction for positive current convention is also reversed in the second diode. In the second diode, positive current flow as represented by (7.12) is from anode to cathode (as for the first diode), which is now, of course, from its output to input. Therefore, the total output port current in Figure 7.9(a) is given by

$$I_0 = I_1 - I_2 \tag{7.13}$$

where I_1 and I_2 are the diode currents from anode to cathode in the two diodes, respectively. This is true irrespective of frequency. Therefore, the even components of the diode current simply cancel each other out, and all even spurious products are rejected in the output port. Of course, the IF component of I_2 will contain a negative term since it results from the product of V_{RF} and V_{LO}, which is opposite in the second diode to that in the first. This gives a double negative in (7.13), causing the IF currents to add in phase at the output port as desired.

When considering spurious components in balanced structures, remember that the spurious signal results from mixing within the nonlinearity of the mixer. Thus, the input signals that cause the spurious are generally in-band, and the input hybrid behaves as designed at the in-band frequencies. However, the spurious components of the output current will fall at many frequencies, and the output hybrid may not behave in the same way as at the fundamental frequency of the desired output signal. Thus, with the structures of Figure 7.12(a), we need to be careful because the rat race itself forms part of the output IF circuit between the diodes and the output port. In the same way as we use an output hybrid, the frequency response of that circuit must be examined to determine whether the frequency components of the diode currents add or subtract at the output *at the frequency we are considering*. However, in this case, the path lengths from the diodes to the IF output port are of equal length, so (7.13) and phase synchronism are maintained across all frequencies, and the cancellation of the even components still occurs.

With the LO injected in the difference port as shown in Figure 7.12(a), the LO appears opposite in sign at each diode input, and the (2,1) spurs are rejected but not the (1,2) spurs. As above, the fundamental LO voltage will be $+V_{LO}$ at each diode anode, while the fundamental RF will be $+V_{RF}$ at one anode and $-V_{RF}$ at the other. Thus (2,1) terms will be of the form $(\pm V_{RF})^2 V_{LO}$, while (1,2) terms will be like $\pm V_{RF}(V_{LO})^2$. Because the (2,1) terms are always positive, they cancel at the output IF port [due to the current subtraction that occurs from (7.13)] while the (1,2) terms are of the opposite sign and will reinforce each other. If the LO and RF connections to the sum and delta ports are reversed, mixing still results but the (1,2) spur is rejected instead of the (2,1).

The characteristic impedance of the rat-race ring itself is 70.7Ω in a 50-Ω system. The ring presents a 50-Ω impedance to each diode at the RF/LO because the quarter-wave sections of the ring behave as two quarter-wave transformers in parallel, that transform 50-Ω loads into 100Ω that then appear in parallel. At the IF frequency, the two diodes appear in parallel with each other, thereby reducing each diode output impedance of perhaps up to 200Ω to a more manageable value.

In general, the VSWR of the single-balanced mixer depends on the balun used to combine the RF and LO signals. Use of a 90° hybrid only provides a good input match if the two diodes have equal reflection

coefficients and the RF and LO have good source impedances at both their own and the other's frequency. With the 180° hybrid, the VSWR depends on how well we can match each diode to 50Ω. Of course, the LO/RF isolation in the 180° coupler is always good independent of the match, because the RF signal is injected at a virtual ground to the differential LO signal, provided again that the diodes have equal behavior. Similarly, there is good LO/IF isolation. Both types of coupler in the single-balanced mixer will reject spurious products involving even m and even n.

7.2.3 Double-balanced mixer

Figure 7.13 shows how four diodes can be used in a double-balanced structure. Double-balanced mixers are usually the mixer of choice because of their superior suppression of spurious mixing products and good isolation between all ports. From the symmetry of the structure, there is a virtual ground to the local oscillator signal across the two terminals *RR'* of the RF balun: The LO voltage is the same at both nodes, so no LO voltage appears across the RF input. Similarly, a virtual RF ground exists across the two nodes *LL'* where the LO balun is connected, since from symmetry the RF voltages at these two nodes is identical. Thus, no RF voltage appears across the LO port.

Intuitively, each side arm consisting of a diode pair is switched on and off alternately by the strong local oscillator signal. The conductance waveform of each side arm is ideally a square wave. If again we use phasor analysis, then when the LO voltage is instantaneously positive, as shown in Figure 7.14(a), the conductance waveform can be represented as a vector rotating at the LO frequency. Diodes D1 and D2 are both instantaneously on and off together so the conductance waveforms are in phase. The RF

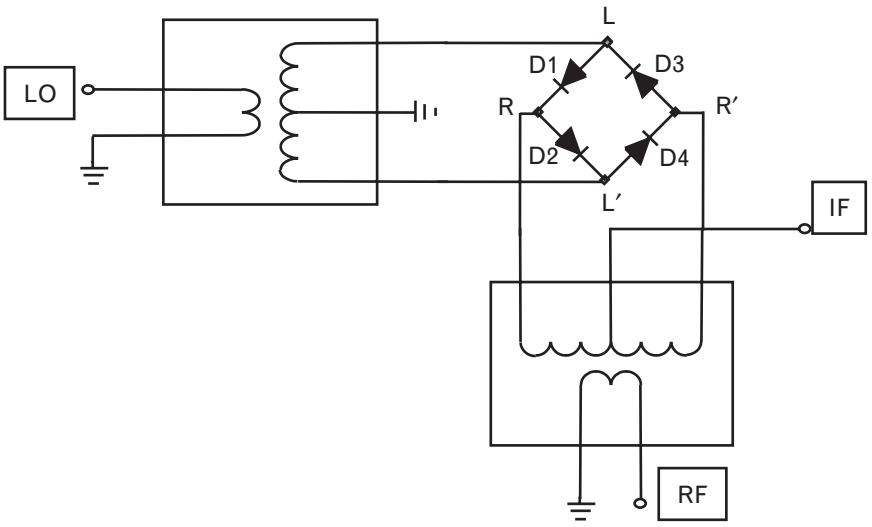


FIGURE 7.13
Circuit topology for a double-balanced mixer using transformer baluns.

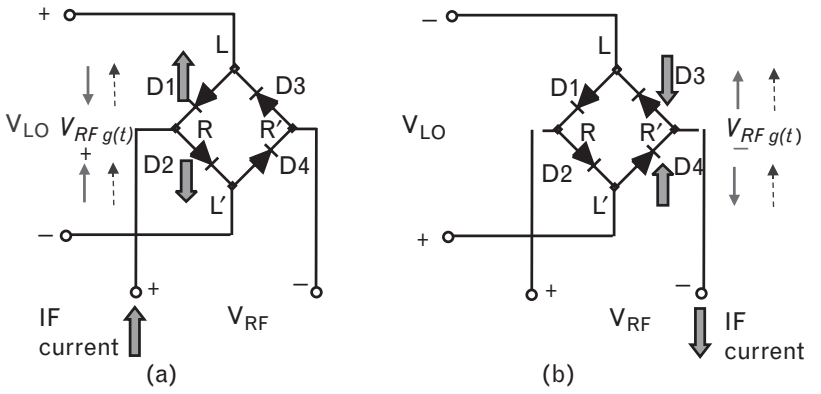


FIGURE 7.14
Phase relationships between the conductance, RF, and IF signals: (a) with LO positive, and (b) with LO negative. (After: [1].)

voltage is applied at the junction of D1 and D2, and is shown instantaneously positive. Because of the reversal of diode D1 with respect to D2, it tends to turn diode D2 harder on, and to turn D1 slightly off. The RF phasors are therefore shown in opposite senses. The difference in phase between the conductance and RF for each diode determines the sense of the IF current, so the IF currents are also in opposite senses, and sum at the junction, flowing *into* the RF lead *R*. The circuit is completed by IF current flowing out of nodes *L* and *L′* and out of the center tap of the LO transformer. Although it appears that current is flowing the wrong way through diode D1, we need to bear in mind that this diode is saturated in a hard on state by the LO, and application of a negative voltage at its anode only slightly modulates the LO current. The positive RF voltage attempts to reduce the forward LO current slightly, so the IF currents shown should be thought of as incremental.

Figure 7.14(b) shows the opposite LO cycle, when diodes D3 and D4 are both switched on. The conductance waveforms are in phase, the RF voltages applied in an opposing sense to the diodes because they are reversed to each other, so the incremental IF current is forced *out of* the junction of D3 and D4 as shown. The IF current flows *out of* the RF lead *R′*. The circuit is completed by IF current flowing into the center tap of the LO transformer and into nodes *L* and *L'*. Thus, the IF current is balanced in the LO transformer across each LO cycle since it flows alternately in and out of the center tap, and no IF signal appears across the LO.

When the RF voltage reverses, a similar analysis can be performed for each half-cycle of LO voltage. The IF current will now flow *out of* node *R* and diodes D1 and D2 when they are switched on by the LO, and *into* node *R'* and diodes D3 and D4 when they are switched on during the other half LO cycle. Thus, the IF current flows in a reverse sense through the RF balun when the RF voltage reverses to the above, and if the center tap is ideal, no IF voltage appears across the RF output.

Some care is needed to determine the diode embedding impedances in a balanced circuit, and not only because of the transformation ratio of the baluns. If the two halves of the center-tapped secondary of each balun each

have the same number of turns as the primary, then for the overall 2:1 transformation ratio the diode quad sees four times the LO impedance (i.e., 200Ω in a 50-Ω system). Then from symmetry, the circuit may be split along the virtual ground across the middle of the quad and modeled as an LO voltage with a source impedance of 100Ω driving a single antiparallel-connected diode pair. To the LO, only one diode of the pair is ever switched on at any time, so each diode sees 100Ω at the LO. When the same split of the quad is made through the RF virtual ground for RF analysis, the antiparallel diode pair now appears instead as two variable conductances in parallel (e.g., D1 and D2). They are switched on and off at the same time, at the LO frequency [each having the conductance given by (7.2)]. The effective RF impedance seen by each diode is thus 200Ω since the RF current from the effective 100-Ω RF source splits in half between the two diodes. Similarly, the embedding impedance for each diode at the IF frequency is also 200Ω.

Because of the symmetry, even-order spurious responses, such as the infamous (2,2) product, are also rejected by the double balanced mixer. Because the RF voltage is split between four diodes, the RF power in each diode is one-quarter that of a single-balanced mixer, so the 1-dB compression point and third-order intercept point are almost 6 dB higher. However, four times as much LO power is now required to pump the diodes to the same degree. The conversion loss is the same, because the RF power is split four ways and the IF power recombined four ways; therefore, the increase in intercept point provides a true increase in dynamic range due to the increase in output compared to a single diode. However, beyond about 10 dBm LO power, the increase in intercept point does not rise as fast as the LO power, because the "on" diodes begin to limit the LO voltage across the "off" diodes, which are in parallel. To each diode, the RF current is indistinguishable from the LO current, and the total RF swing is therefore limited in the "off" condition. This can be improved by using two or more diodes in series in place of the single diodes shown.

Double-balanced mixers can be purchased ready-made for integration. An example is the M/A-COM EMD40-2400L, whose data sheet in shown in part in Figure 7.15. The package employs a quad-ring of diodes, with internal RF and LO baluns, in a surface mount SO-8 package. It is usable over the 1,400- to 2,500-MHz frequency range. The RF and LO signals are applied to the baluns single-ended, and the IF is also single ended. Such packages are ideal for low-power applications such as hand-held radios. This mixer has a particularly low LO power requirement for a double-balanced mixer (+3 to +7 dBm), a conversion loss and noise figure better than 7 dB, and typically 35 dB of RF-LO isolation, which is determined completely by the balance of the RF and LO transformers. With +7 dBm of input LO power, the input third-order intercept point is a minimum of 8.5 dBm from 1,700 to 2,000 MHz (typical is 11 dBm), and the input 1-dB compression point a minimum of 1 dBm.

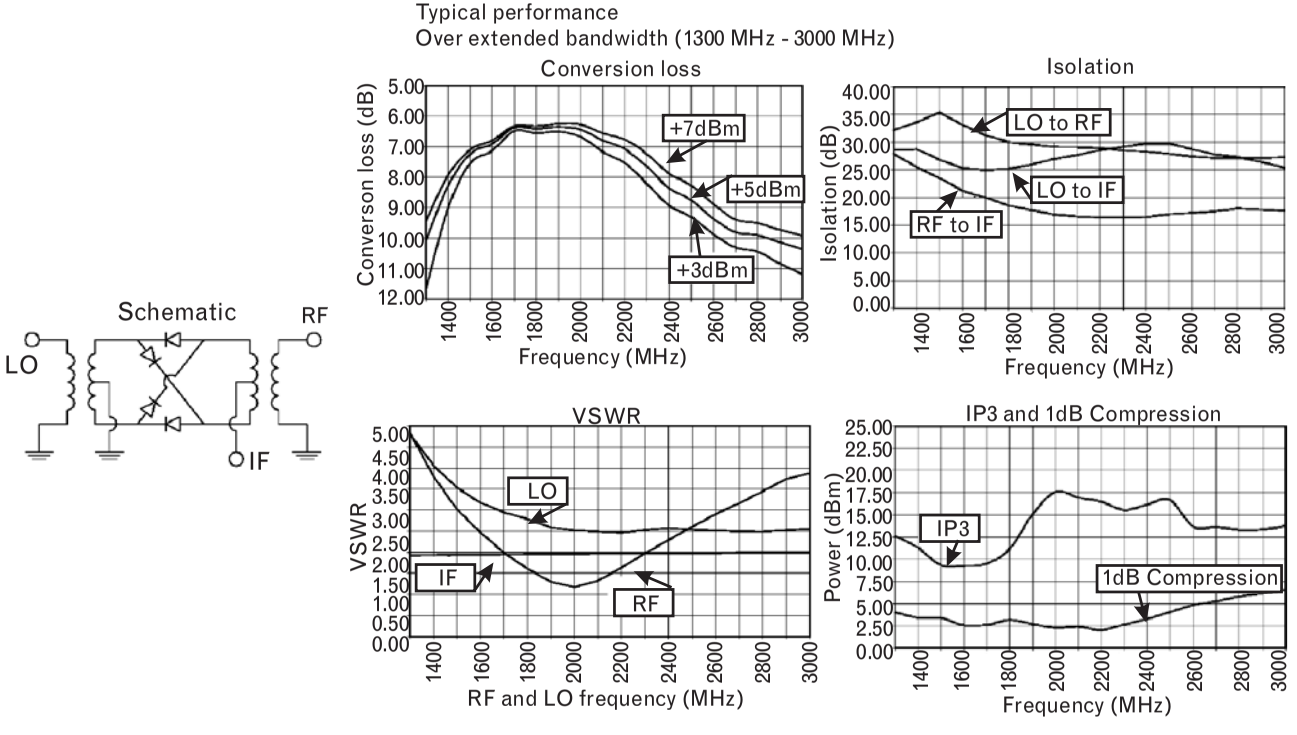


FIGURE 7.15 *EMD40-2400L double-balanced mixer schematic. (Courtesy M/A-COM, Inc.)*

Double double-balanced mixers, as shown in Figure 7.16, use two double-balanced mixers and a separate IF balun to extract the IF. This further extends the dynamic range of the mixer, since the RF power is now shared between eight diodes, although the LO input power level must be again increased to drive the diodes sufficiently hard. However, the use of an external balun now permits the IF frequency range to be well separated from either the RF or LO in frequency. Since center taps in the RF or LO transformers are no longer required to feed the IF current flow, isolation between the RF, LO and IF ports is almost complete, limited only by the degree of mismatch between the diodes themselves.

7.2.4 The image problem in mixers

In Volume I, Chapter 3, and previous sections, we discuss at some length how any signal at the image frequency, either noise or an interfering tone, will mix and produce an unwanted response at the intermediate frequency. Even if there is a null signal at that frequency, any noise there will add to the IF noise floor and reduce the system dynamic range.

There are a number of ways to solve the image problem:

1. Use a homodyne system in which the LO frequency equals the RF, so that the IF is centered at dc and the image frequency is the same as the signal.
2. Use a double-sideband system, in which the image frequency is occupied by either the upper or lower desired sideband. This is not so common since it is wasteful of spectrum, and only solves the

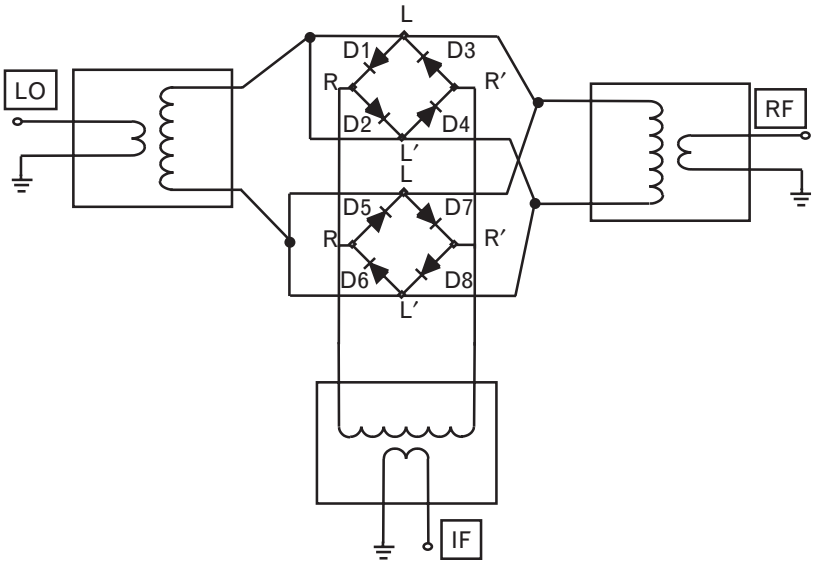


FIGURE 7.16
Double double-balanced mixer.

image problem for the final downconversion if there are multiple conversion steps in a radio.

3. Prevent the image frequency from entering the mixer by bandpass filtering around the RF frequency on the RF input port, as shown in Figure 7.17. Such a mixer is known as an *image enhanced* or *image recovery* mixer and needs to be purpose built for the particular application. Such a system is only effective when the IF is relatively high, so that the skirts of the bandpass filter can fall rapidly enough to eliminate the image. It can be even more difficult in mixers in which the RF and LO signal are fed into the same port, such as in a single-ended mixer, for then any filtering will have to be broadband enough to include both the local oscillator and RF. Short-circuit termination of the image by the filter is usually preferred.

7.2.4.1 The image-reject mixer and quadrature upconverter

The solution most commonly adopted to solve the image problem, at least when integrated circuits are used to implement a system, is to use an image-reject mixer. Although a fairly complicated structure, the image-reject mixer is seen frequently in radio systems because it is also the core of a single-sideband modulator. The principle is shown in Figure 7.18.

A 90° hybrid is used to split the LO signal that pumps the two balanced mixers, each fed in-phase with the RF. The resulting IF outputs are combined in a second 90° hybrid (at the IF frequency) so that one of the IF combinations contains only the signal frequency, and the other contains only the IF response from the image. In this way, the IF image power can be dissipated to yield a 3-dB sensitivity improvement in the IF signal port.

Figure 7.18 shows a signal of the form $\cos(\omega_S t)$ entering the RF input, where $\omega_S = 2\pi f_S$ and f_S is the desired signal frequency. Suppose the LO signal has the form $\cos(\omega_{LO} t)$ and $\cos(\omega_{LO} t - 90)$. If we consider only phase differences rather than the absolute phase as we progress through the system, the output signal from the top mixer is of the form $\cos((\omega_S - \omega_{LO})t)$ while that from the bottom mixer $\cos((\omega_S - \omega_{LO})t + 90)$. Now if we take the case where the mixer is a low-side downconverter where $\omega_S > \omega_{LO}$ (and this can be done without loss of generality), then the top input to the output IF

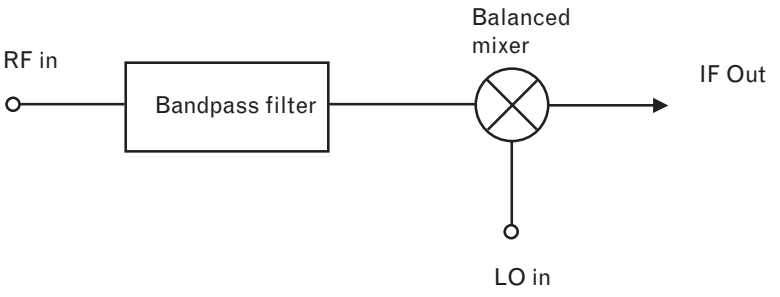


FIGURE 7.17
Principle of an image enhanced mixer.

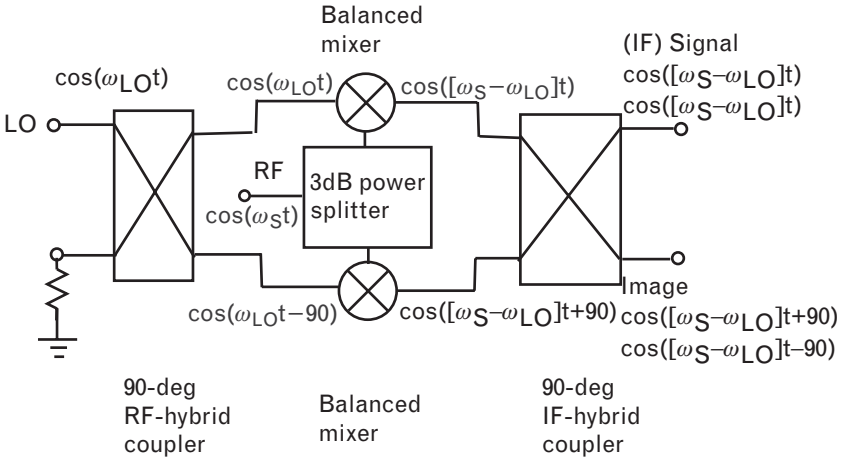


FIGURE 7.18
An image-reject mixer, in which the RF input and LO are mixed to produce an IF resulting from the signal frequency and an IF from the image frequency. The path of the desired signal through the mixer is shown, and assumes $\omega_S > \omega_{LO}$.

coupler is just $\cos(\omega_{IF}t)$ and the bottom input is $\cos(\omega_{IF}t + 90)$. Considering first the top input, it couples to the top and bottom output ports of the IF coupler as $\cos(\omega_{IF}t)$ and $\cos(\omega_{IF}t - 90)$, respectively, because the bottom arm introduces a phase lag of 90°. The bottom input couples to the top port as $\cos(\omega_{IF}t)$ because it undergoes the 90° phase delay, and $\cos(\omega_{IF}t + 90)$ to the bottom port. Thus, the two signals cancel at the lower port of the output because they differ in phase by 180° and add in phase at the upper (signal) port.

Figure 7.19 shows the path of the image signal through the same mixer. The output from the top balanced mixer is of the form $\cos((\omega_I - \omega_{LO})t)$, and from the bottom mixer $\cos((\omega_I - \omega_{LO})t + 90)$ analogous to before where ω_I is now the image frequency. However, $(\omega_I - \omega_{LO})$ is now a negative frequency since $\omega_I < \omega_{LO}$, so using

$$\cos(-\omega t) = \cos(\omega t)$$

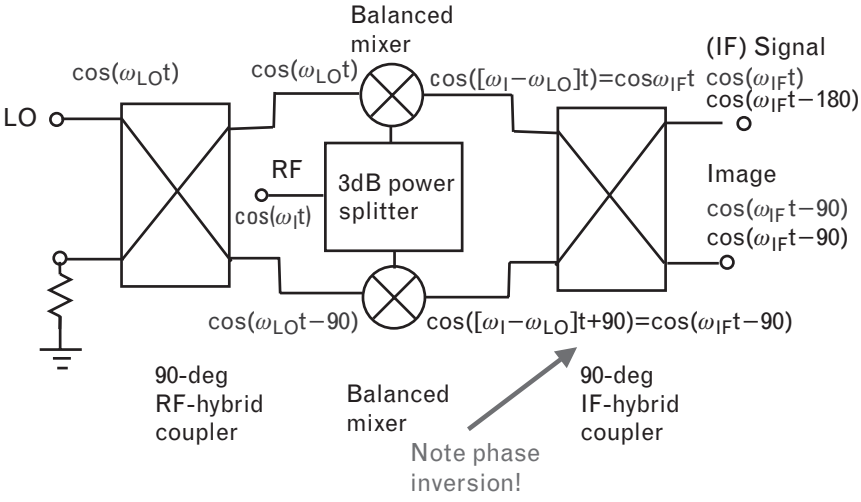


FIGURE 7.19
The image-reject mixer showing the path of the unwanted image signal through the mixer. This assumes, consistent with before, that $\omega_I < \omega_{LO}$.

these can be written as positive frequencies $\cos(\omega_{IF}t)$ and $\cos(\omega_{IF}t - 90)$, respectively, where $\omega_{IF} = \omega_{LO} - \omega_I$ is now positive. The output from the top mixer produces outputs from the IF hybrid coupler of the form $\cos(\omega_{IF}t)$ and $\cos(\omega_{IF}t - 90)$, respectively. The output from the bottom mixer produces an output of $\cos(\omega_{IF}t - 180)$ from the coupler's upper output port, because it is delayed 90°, while the output signal at the lower port is $\cos(\omega_{IF}t - 90)$. In this case, the IF frequencies resulting from any signal at the image frequency cancel at the upper port because they are 180° out of phase and sum at the bottom (image) port where they are in phase.

As a consequence, the signal is directed to one port and contains no translated (or folded) component from the image frequency. The image component is instead directed to a separate port, where it can be dissipated in a 50-Ω termination.

The same principle works if the 90° phase shift is introduced into the RF path instead of the LO, although this can generally introduce undesired loss, hence noise. However, one 90° hybrid can be avoided if the in-phase and quadrature RF signals are already available, as for instance, from an earlier quadrature downconversion. Alternatively, we note that the quadrature LO output $\cos(\omega_{LO}t - 90)$ is just $\sin \omega_{LO}t$, so the LO hybrid coupler can be eliminated if the LO sine and cosine are already available. Such a quadrature LO signal is often already available in many integrated circuits from the phase-lock loop, or by halving the frequency during the generation of the LO signal.

In digital superheterodyne systems, such as some of those examined in the next chapter, the output IF hybrid coupler of what otherwise appears to be an image-reject mixer may apparently be omitted. In that case, the image could have already been rejected by earlier RF filtering or the choice of an appropriate IF. Then, the mixer is simply a quadrature mixer to purely generate quadrature channels for I and Q processing of the amplitude and phase modulation, rather than for image rejection. Also, the quadrature IF signals at the mixer outputs can later be digitally mixed to baseband, and the quadrature function for image separation is then performed digitally. This requires high dynamic-range ADCs since both the signal and image are still present at the sampling output of the analog stage.

Mathematically, the image-reject mixer operation can also be explained as a form of complex mixing. We recognize the local oscillator and its quadrature component as a complex signal

$$\cos(\omega_{LO}t) - j\sin(\omega_{LO}t) = e^{-j\omega_{LO}t} \tag{7.14}$$

which in the frequency domain has only a single negative frequency component at $-\omega_{LO}$. This is illustrated in Figure 7.20. When this is multiplied by the RF signal at $\pm\omega_S$, a convolution occurs in the frequency domain, and the RF component is shifted to a frequency $\omega_S - \omega_{LO}$, which is the IF,

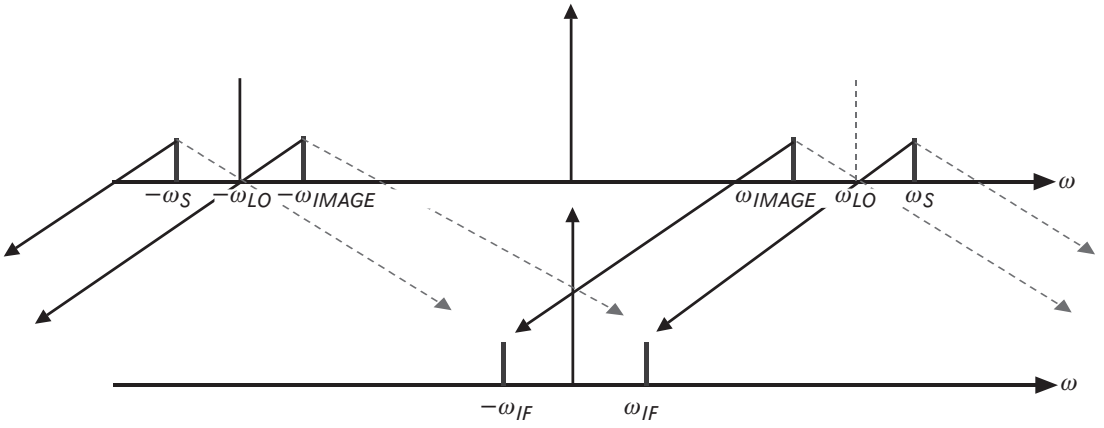


FIGURE 7.20 *The frequency-shifting properties of the image-reject mixer. A quadrature LO is equivalent to a complex representation of a single frequency component only. The solid lines show the resulting frequency translation in an image-reject mixer; the dashed lines are additional translations that occur in a normal balanced mixer.*

and to $-\omega_S - \omega_{LO}$, which is out of band. The single LO also shifts the image at $\omega_S - 2\omega_{IF}$ to $\omega_S - 2\omega_{IF} - \omega_{LO} = -\omega_{IF}$ (and likewise the negative image at $-\omega_S + 2\omega_{IF}$ shifts out of band). The output image IF (which is negative) and the signal IF can then be separated in the output IF hybrid, since the two components are no longer aliased onto each other.

A single mixer, of course, multiplies by $\cos(\omega_{LO}t) = 1/2(e^{-j\omega LO t} + e^{+j\omega LO t})$ which has two frequency components, one at $-\omega_{LO}$ and the other at $+\omega_{LO}$. As shown by the dotted lines in Figure 7.20, it is the second of these components at ω_{LO} that normally causes the negative image at $-\omega_S + 2\omega_{IF}$ to upconvert to the same positive IF frequency ω_{IF} as the signal, where it can no longer be isolated from it.

When the image-reject mixer is used in the reverse direction, it functions as an upconverter, and becomes an I-Q modulator. The signal flow above is simply reversed, and the I and Q baseband channels of a phase-modulated system provide the equivalent to the two output 90° phase-shifted IF signals, each directly feeding one of the two component mixers. The sine and cosine of the LO again drive the mixers, and the modulated RF output can be simply summed in-phase at the output of the two mixers.

An analog *single-sideband* (SSB) modulator functions the same way, but uses a 90° coupler at the baseband input to split the modulation signal; the upper and lower sideband of the RF are selected from either the difference or sum port of a 180° hybrid at the output of the two mixers, again driven by a quadrature LO. The baseband spectrum is simply linearly translated to RF as a single-sideband.

The structure of the image-reject mixer and the single-sideband modulator are therefore similar: the former typically uses a quadrature shift

in one of the inputs (RF or LO) and one at the output IF, as in Figure 7.18; the latter typically uses a quadrature phase shift in both input and the LO, and sums (or subtracts) the output from the two mixers directly. In both cases, maintenance of amplitude balance and phase quadrature is crucial to rejection of the unwanted sideband and the LO (carrier) itself. If K is the linear amplitude imbalance between the I and Q channels and Φ the phase deviation from quadrature, the *residual sideband suppression* (RSB) in dBc below the desired sideband is given by

$$RSB = 20\log_{10}\sqrt{\frac{K^2 - 2K\cos\Phi + 1}{K^2 + 2K\cos\Phi + 1}} \tag{7.15}$$

Typical levels of RSB in off-the-shelf quadrature upconverters are –40 dBc at 895 MHz. Carrier suppression is of the same order.

7.2.5 Harmonic components in mixers

7.2.5.1 Subharmonic mixers

At very high frequencies, into the millimeter-wave region, it can become increasingly difficult to generate a LO-signal strong enough to downconvert the RF into a reasonable IF. It can also become increasingly difficult to build a low phase-noise oscillator source. In such a case, a harmonic mixer can be used. This mixer drives the LO at a low fundamental frequency, to generate a square wave conductance waveform, but uses a higher harmonic of the LO waveform to mix with the fundamental of the RF. In effect, we select the "spurious" response $m\omega_{RF} \pm n\omega_{LO}$ with $m = 1$ and $n > 1$ as the intermediate frequency. If a single diode is used, n will be chosen odd because these are the strongest components in the square-wave conductance waveform.

One implementation of a subharmonic mixer is shown in Figure 7.21. The RF frequency is, for instance, in the millimeter-wave region, while the LO and IF will be in the microwave or RF regions and can share a common filter. This is not necessary, however, and the LO could be applied through a separate filter across the same terminals of the diode antiparallel pair. Depending on the frequencies of each, the structure of the subharmonic mixer can simplify the problem of isolating the RF and LO compared with a fundamental frequency mixer.

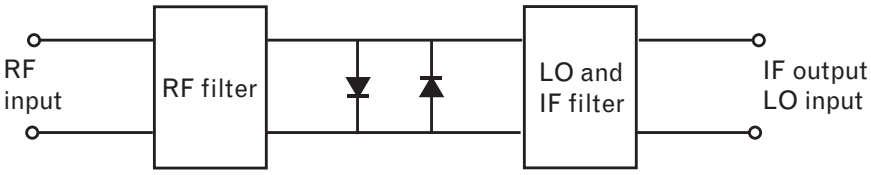


FIGURE 7.21
A subharmonic mixer.

In the mixer shown, an antiparallel diode pair is used, so that each diode conducts on alternate half cycles of the LO, 180° apart. Since the RF is applied to both diodes in phase, the fundamental IF currents will be 180° out of phase and the (1,1) response will cancel since the output current is just the two diode currents in parallel. Since the pair conducts twice every fundamental cycle of the LO, the dominant response will occur for $n = 2$ although higher even-harmonics of the LO can also be used. The conversion loss of such mixers is quite poor and will rarely be better than 10 dB. However, as suggested above, they are useful at high microwave frequencies or for multiband applications that can share a single mixer and LO structure. Applications of the latter include car radar detectors, where the RF radar frequencies can be either in X- or Ku-band, or in multiband cellular phones that operate in both the 900- and 1,800-MHz bands.

Any input frequency around each harmonic of the LO frequency will alias down to the IF. Although the input RF noise will be filtered by the input bandwidth of the mixer, any noise generated by the switch cannot be avoided. In addition, in a harmonic mixer the RF bandwidth is intentionally quite high and there may be several harmonics of the LO that will alias noise onto the IF. The noise performance of such mixers, therefore, deteriorates as a function of the harmonic frequency chosen, because of the increasing noise frequencies that are downconverted.

FETs can also be used as subharmonic mixers. One topology implemented in integrated circuits is to short the drain and source together and to use two FETs as diodes in an antiparallel diode-pair combination. Another option is to drive two FETs in their resistive region, using the fundamental FET resistive mixer detailed below as a building block. Such devices are driven at the gate by the LO while the RF and IF are extracted from the drain, which has no applied bias voltage. Two such class-B devices can then be driven push-pull to create an overall conductance waveform that is rich in even harmonics, to achieve subharmonic mixing on an even harmonic of the LO. The balanced structure automatically eliminates LO leakage at the output port, as the drains of the two devices are simply connected in parallel. Further discussion is beyond the scope of this book; the interested reader is referred to [2] for a more complete description.

7.2.5.2 Mixer spurious products

As in amplifiers, spurious responses in a mixer can also be filtered. We see in Volume I, Chapter 3, that the (2,2) product is frequently a problem in radios, in that the second-harmonic of an unfortunately located interfering signal in-band can often produce a spurious response at the IF through mixing with the second-harmonic of the local oscillator. This is somewhat analogous to the problem of third-order intermodulation distortion in power amplifiers, where two in-band interfering signals can mix through

the third-order nonlinearity of the amplifier and produce an unwanted (2,1) response at the desired RF frequency. In fact, all (*m*,*n*) products in a mixer are problematic, and mixer vendors frequently provide tables showing the relative amplitudes of each response under given LO drive conditions.

One way to reduce such products is to short-circuit the higher harmonics of the LO at the intrinsic mixer terminals to lower the power in such responses. An example of this is provided in [3], where short circuits of the second harmonic and third harmonic of the local oscillator signal are placed at the input terminals of a diode. Figure 7.22 shows the resultant decrease in the output power of the (2,2) and (3,3) spurious responses. In the figure, these responses are referred to as the second and third harmonics of the IF since the same RF signal is used for both the IF and intermodulation measurement. In a radio, of course, it is an RF interfering signal at a slightly shifted frequency whose second- or third-order response falls onto the *same* IF that causes the problem. As expected, reducing the second or

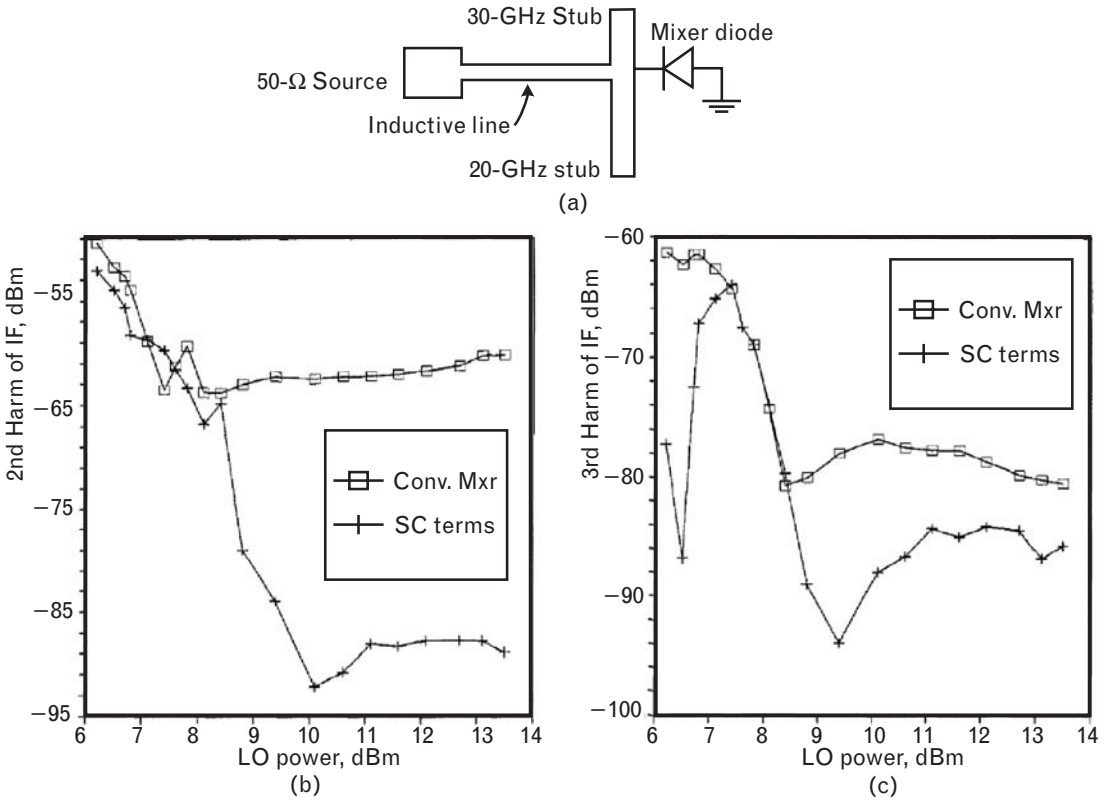


FIGURE 7.22 *(a) A single-ended mixer with short-circuited second and third-harmonic LO terminations at the input. (b) The (2,2) spurious response and (c) the (3,3) spurious response, without ("conventional mixer") and with ("short-circuit terms") the harmonic terminations. (From: [3]. © 1992 IEEE. Used with permission.)*

third harmonic of the local oscillator reduces its harmonic products by 20 to 25 dB and 10 to 15 dB, respectively.

Figure 7.23 shows the conversion loss of the mixer with and without the short-circuited harmonic terminations, as a function of LO drive. As expected, the conversion loss becomes less as the LO power is increased, because the mixers are pumped more efficiently and the conversion coefficient g_1 from the conductance increases as the conductance waveform becomes more square. However, adding the short-circuited terminations makes the conversion loss worse by up to 1 dB. This may not be severe for some applications, but it illustrates why harmonic tuning in mixers is not prevalent. It occurs because the harmonic terminations in Figure 7.22(a) prevent second and third-harmonic LO voltages being impressed across the diode. This causes the conductance waveform to be sinusoidal rather than square as in (7.2), and thus the conversion coefficient g_1 to be reduced. To minimize the distortion ratio, it is desirable in a mixer that the conductance waveform in (7.1) contains a fundamental component as large as possible, and harmonic components as small as possible. However, this cannot be achieved practically because the fundamental component of a square wave, with all its harmonics, is larger than that of a sinusoid with the same peak-to-peak swing between on and off states. However, the possibility of achieving a reduction in the (2,2) product without affecting g_1 is real, given

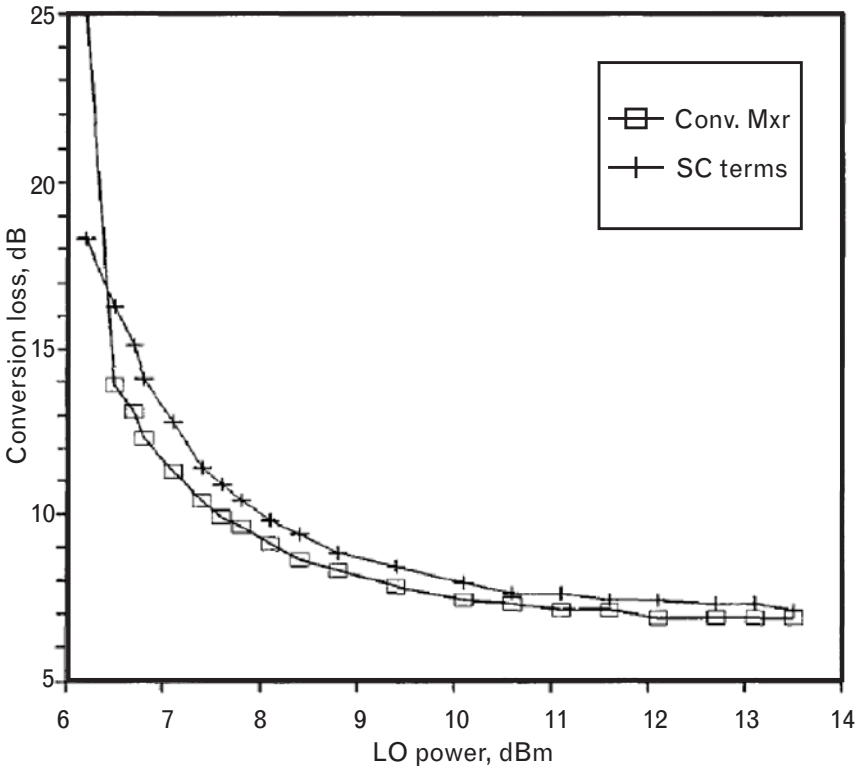


FIGURE 7.23
The conversion loss of the single-ended mixer with and without short-circuited second and third-harmonic terminations to the LO. (From: [3]. © 1992 IEEE. Used with permission.)

an ideal square wave conductance waveform contains no even harmonics. Since this is already achieved with the balanced mixer structures, this has not received any practical attention.

7.3 Transistor mixer design

The possibility of using a transistor with its three terminals compared with a diode's two seems to hold the promise that the RF, LO, and IF ports can now be independent. It is therefore somewhat ironic that the most common active transistor mixer combines the RF and LO signals externally in a balun, in the same way as for a diode mixer.

The principal reason for doing this is that any signal applied to the input terminal (base or gate) of the transistor can be amplified. A mixer where the transistor is biased to provide transconductance, and possibly amplification, is known as an *active mixer*. The bipolar transistor, FET, HEMT, and dual-gate FET can all be used as active mixers. Sometimes the device is not biased at all and is used as a variable resistor. The channel of the FET can serve this useful purpose, and such a device is then known as a *resistive mixer*. We will examine active and resistive transistor mixers separately in the following sections.

7.3.1 Active transistor mixers

As their name implies, active transistor mixers are able, through their applied bias, to provide conversion gain. The conductance waveform (7.1) is now generated by the transconductance of the device, so the local oscillator signal is always applied at the transistor's base or gate to generate the switched transconductance. The RF voltage must then be applied at the input as well, and the resulting IF current is always taken from the output of the transistor. Conversion gains of 10 dB or higher can be achieved, and output third-order intercept points around 15 to 20 dBm attained with very moderate LO power requirements (frequently as low as 0 dBm).

However, many of the same issues exist as for a single-ended mixer, such as the balun problem. Both the RF and LO voltage are applied to the same port—either the base or the gate—so separate filtering is required to isolate the two. As for single-ended diode mixers, active transistor mixers can be used in balanced structures to improve the even-order spurious response and to obtain isolation between the RF, LO, and IF ports.

When used in radio systems, the output intercept point of mixers is frequently the determining constraint on the spurious-free dynamic range. If the mixer has gain, the intercept point referred to the input is lower than at the output, so that too much gain in a mixer is not necessarily a desirable property as it can undesirably lower the dynamic range.

Figure 7.24 shows the basic principles in matching and biasing any active transistor mixer. As stated, the RF and LO are externally filtered and combined to drive either the base or gate of the device. The IF current is taken from the collector or drain through an IF filter. The device is biased with the transfer characteristic shown, similar to that for a class-B amplifier. Some transistor mixers have also been biased in the transistor saturation region of the load line, near the knee of the I-V curves, to similar effect. However, because the dc efficiency is much worse when biased at a high quiescent current point such as this, and because their properties are no better than for devices biased at class-B, they are not popular.

As the local oscillator drives the base or gate, the device swings on along the load line only during positive-going voltage swings. The classical half-wave rectified sinusoidal current results. If the device is an ideal device, with constant g_m, the transconductance waveform will be a square wave, at the local oscillator rate. For real devices, where g_m is small as the device begins to turn on, the ideal square wave transconductance will have a finite rise time and look more sinusoidal during the on cycle. For ideal bipolar transistors, where g_m is linearly related to the collector current, the transconductance will be exactly sinusoidal during the on cycle. For non-ideal devices, where g_m is zero at turn-on and maximum at the peak of the current swing, the transconductance waveform will appear as a "pinched" half-sinusoidal waveform with the peaks of each sinusoid "squeezed"

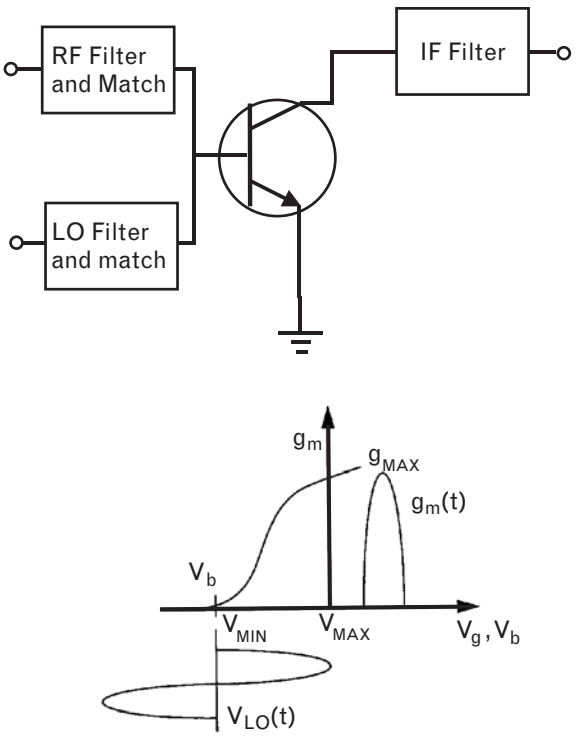


FIGURE 7.24
Basic principle for the design of an active transistor mixer.

together. In all cases though, the transconductance will have a dominant fundamental component to it, so that we may write similar to (7.1)

$$g(t) = g_0 + g_1 \cos(\omega_{LO} t) + \ldots$$
$$g_1 = k g_{MAX} \tag{7.16}$$

where k is a constant around 0.5 (for a sinusoidal half-wave rectified g_m) or $2/\pi$ (for a square wave g_m), and g_{MAX} is the peak value of g_m when the transconductance (and generally current) is maximum.

As the small-signal RF voltage is applied to the base or gate, the output current takes the form of (7.3). The component of the collector or drain current at the IF is selected by an output IF filter, which should attempt to short-circuit all other frequency components in the current. Some care is needed in this respect, since filters often present reactive terminations—rather than short circuits—outside their passband. In particular, all RF and LO output currents should be short-circuited by this filter at the fundamental and its harmonics, to prevent any collector voltage at these frequencies that could potentially feedback across the collector-base or drain-gate junctions into the base or gate and remix within the device. This will also hold the dc collector or drain voltage constant and keep the device in its active region during the on cycle. The IF filter can also attempt to match to the IF output, although if the IF frequency is too low, some caution is needed because the transistor gain will be high and the device may be unstable. Matching will also be more difficult because the device output will appear as a very high impedance current source loaded by the output capacitance. In such a situation it may be more prudent to forgo matching the IF output and sacrifice the gain for stability.

Similarly, the input filters, which are necessary to achieve RF to LO isolation and prevent radiation of the LO back through the antenna or other RF input, should attempt to short-circuit all unwanted frequencies (i.e., those other than the RF and LO) so there are no interfering voltages appearing at the input. The input should be matched to the RF to maximize conversion gain and noise figure, and if possible, to the LO as well for LO power transfer. In particular, the image frequency should be short-circuited (if possible), as well as the IF, so neither noise nor spurious signals are amplified by the device. It is important that the device not behave as an amplifier at the IF, especially if the IF is low where the device gain is high. As a general rule in mixer design, all undesired frequencies should be short-circuited at *both* the input and the output to minimize distortion, noise, and for stability. In the following circuits, these terminations are represented by parallel L-C circuits connected in shunt at the input and output, tuned to resonate at the desired frequency.

Transistor mixers are not unilateral devices, and the output termination can significantly affect the input RF match, which is generally critical. One

way to deal with the interaction of input and output is to simulate the quasi-linear *S*-parameters of the device, using the RF frequency at the input and the IF frequency at the output. Then, we may use for the mixer exactly the same bilateral conjugate matching techniques as we developed for the small signal amplifier. In this case, the *S*-parameters must be simulated, since they are parameters relating a different frequency at the output to the input, and incremental in that the local oscillator sets the operating point of the device about which the RF and IF are incrementally applied [4].

Many of the advantages of balanced structures, such as improved isolation, reduced spurious response, and improved intercept point, can be achieved for active mixers in the same way as for diode mixers. Similar principles apply, and they are discussed more fully in the section on resistive FET mixers.

7.3.1.1 Active bipolar transistor mixers

Figure 7.25 shows the basic principles of an active mixer using a bipolar transistor, where we have transformed the RF and LO to show the equivalent circuit after the effect of matching. The RF and LO voltages are assumed to be combined externally, through some external summing network. At low frequencies, this could be a simple resistive adder; at microwave frequencies it could be two microstrip directional couplers coupled to the base input transmission line. For the purposes of analysis, we represent the RF or LO by a single voltage source and assume that the input matching network converts its 50–Ω source impedance into some resistor *R* and inductance *L* seen from the base of the transistor. If we assume the input is conjugately matched, and the RF and LO frequencies are higher than the 3-dB roll-off frequency of the transistor, then

$$\begin{aligned} R &\approx r_b \\ X_L &\approx -X_{C_\pi} \end{aligned} \tag{7.17}$$

where, from Chapter 3, the transistor is represented simply by its base resistance r_b and the diode equivalent circuit r_π and C_π. At this frequency, the device input reactance X_C is approximately that of C_π alone. We have neglected the feedback capacitance C_μ. This can cause a high Miller capacitance and degrade the frequency response of the mixer, so that if very high frequency operation is required a common-base structure could be considered instead, although then the mixer conversion gain will be lower. As noted earlier, the RF, LO, and IF filters are shown as shunt L-C circuits tuned to these frequencies, so that all unwanted frequency components are short-circuited at their respective ports.

Now consider Figure 7.25 from the perspective of the applied large-signal LO voltage. The intrinsic base-emitter voltage v_{IN} must swing

between some values V_{MAX} and V_{MIN} so the device achieves the class-B conductance waveform of (7.16). If the zero-to-peak magnitude of the applied LO voltage (transformed through the input matching network) is $|V_{LO}|$, then the current that results in the base is

$$I_{LO}(t) \frac{|V_{LO}|\cos(\omega_{LO}t)}{2r_b} \tag{7.18}$$

where we have assumed conjugate matching so $R = r_b$ and r_π is negligible compared with the reactance of C_π. We may then calculate the intrinsic base voltage as

$$v_{IN}(t) = \frac{I_{LO}(t)}{\omega_{LO}C_\pi} = \frac{|V_{LO}|\cos(\omega_{LO}t)}{2\omega_{LO}C_\pi r_b} \tag{7.19}$$

However, the maximum power available from the local oscillator, into the matched load, is just

$$P_{LO} = \frac{|V_{LO}|^2}{8R} = \frac{|V_{LO}|^2}{8r_b} \tag{7.20}$$

so that using V_{LO} from (7.19), we obtain

$$\begin{aligned} P_{LO} &= \frac{\left(|v_{IN}(t)|2\omega_{LO}C_\pi r_b\right)^2}{8r_b} \\ &= \frac{1}{2}r_b\left(\omega_{LO}C_\pi\right)^2\left(V_{MAX} - V_{MIN}\right)^2 \end{aligned} \tag{7.21}$$

Since the input LO voltage must drive the device class-B between an off condition and some maximum current I_p, the zero-to-peak voltage

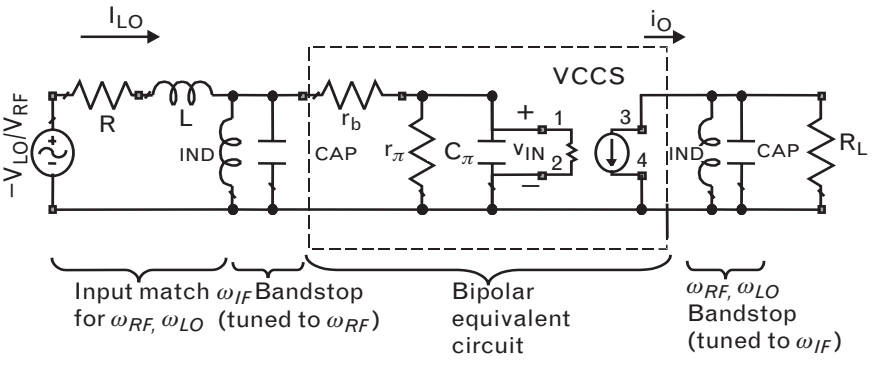


FIGURE 7.25
The small-signal equivalent circuit of an active bipolar transistor mixer.

swing at the base can be approximated by $V_{MAX} - V_{MIN} = I_P/g_0$ since g_0 is the average value of g_m. The voltage at the base will swing an equal amount in the opposite direction, when the device is off.

The conversion gain of the active bipolar mixer can also be calculated by now applying power at the RF and calculating the output IF power. Using (7.16) and (7.19) at the RF, we may write the output current as

$$i_0(t) = g_m(t)v_{IN}(t) \approx \left(g_0 + kg_{MAX}\cos(\omega_{LO}t)\right)\frac{|V_{RF}|\cos(\omega_{RF}t)}{2\omega_{RF}C_\pi r_b} \tag{7.22}$$

which has an IF component of

$$i_{IF}(t) = kg_{MAX}\frac{|V_{RF}|}{4\omega_{RF}C_\pi r_b}\cos(\omega_{LO} - \omega_{RF})t \tag{7.23}$$

The IF output power into a load resistor R_L is then just

$$P_{IF} = \frac{1}{2}|i_{IF}(t)|^2 R_L = \left(kg_{MAX}|V_{RF}(t)|\right)^2 \frac{R_L}{32(\omega_{RF}C_\pi r_b)^2} \tag{7.24}$$

The available power from the RF source is the equivalent of (7.20),

$$P_{RF} = \frac{|V_{RF}|^2}{8R} = \frac{|V_{RF}|^2}{8r_b} \tag{7.25}$$

so from the above two equations, the mixer conversion gain can be determined from the ratio of IF to RF power,

$$G_T = \frac{1}{4}\left(\frac{kg_{MAX}}{\omega_{RF}C_\pi}\right)^2 \frac{R_L}{r_b} \tag{7.26}$$

This equation assumes that the LO power is sufficiently high to drive the device to the peak value of g_m. As the LO voltage swing is reduced, so too is the conversion gain, because the corresponding value of g_{MAX} in (7.26) is reduced proportionally.

Active bipolar mixers have their limitations. The expression for gain varies as $(f_T/f_{RF})^2$, which is perhaps to be expected, since we saw in Chapter 3 that f_T is just the gain-bandwidth product of the transistor. Because they are a minority carrier device, bipolar transistors do not switch well above $f_T/10$ where they will have degraded third-order intercept point and noise figure. In addition, because in large-signal operation the relationship

between the input voltage and output current is exponential, the response to even the small-signal RF is quite nonlinear. Nevertheless, in applications where power is at a premium, such as in pocket pagers, the bipolar mixer is commonly used since the transistor behaves like a single diode mixer with gain. The third-order intercept can be as much as 9 dB above the LO power level, and the noise figure is around 7 dB.

Bipolar transistor mixer example

Let us design a mixer for the 850- to 950-MHz wireless frequency band, with a 45-MHz IF. Assume a low-side LO at 855 MHz so the RF is centered around 900 MHz. In this example we will again use the Infineon BFP640 in the SOT343 package, the versatile HBT first introduced in Chapter 1.

The first step is to match the input of the HBT around the LO and RF frequencies, at 900 MHz. Figure 7.26 shows the input and output *S*-parameters for this device from 10 to 4,000 MHz, when biased with a collector voltage of 2V and 13-mA dc current. It is apparent that the input can be matched to 50Ω using a single shunt inductor at the base, since the device input happens to cross the 50-Ω conductance circle around 900 MHz. The required normalized shunt inductive susceptance to cancel the device capacitance is about $-j0.65$ siemens.

However, an additional requirement is that the input matching network should short-circuit the IF frequency at the input. Therefore, rather

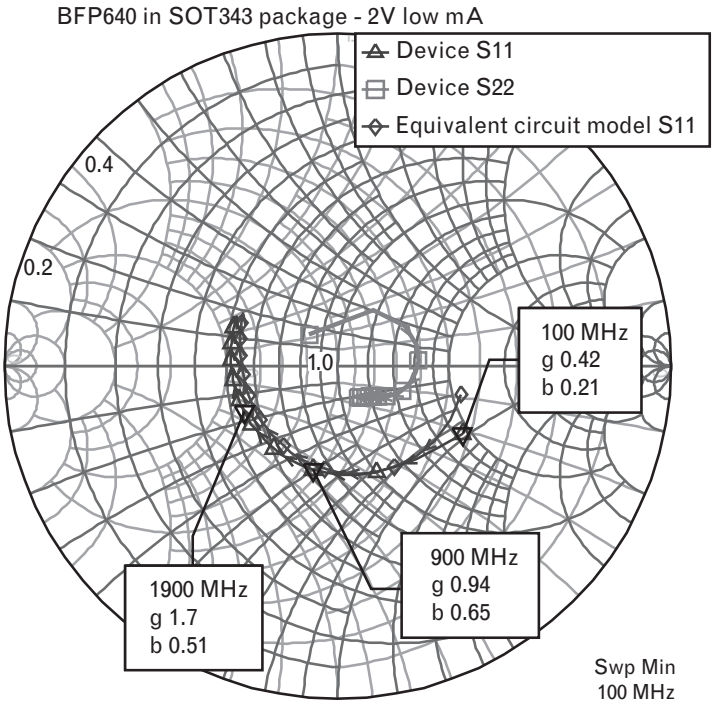


FIGURE 7.26
Small-signal s_{11} and s_{22} of the BFP640 HBT device in its SOT343 package, measured up to 4,000 MHz, at 2V, 13-mA bias. The s_{11} of the transistor input small-signal model (in Figure 7.28) is also shown.

than using just a shunt inductor at 900 MHz to cancel out the input capacitance of the transistor, we will use instead a series-L-C circuit connected in shunt, tuned to resonate at 45 MHz to provide the IF short circuit, and tuned to provide the necessary net susceptance of $-j0.65$ at 900 MHz. The resulting component values for the input match, and the match achieved, are shown in Figure 7.27.

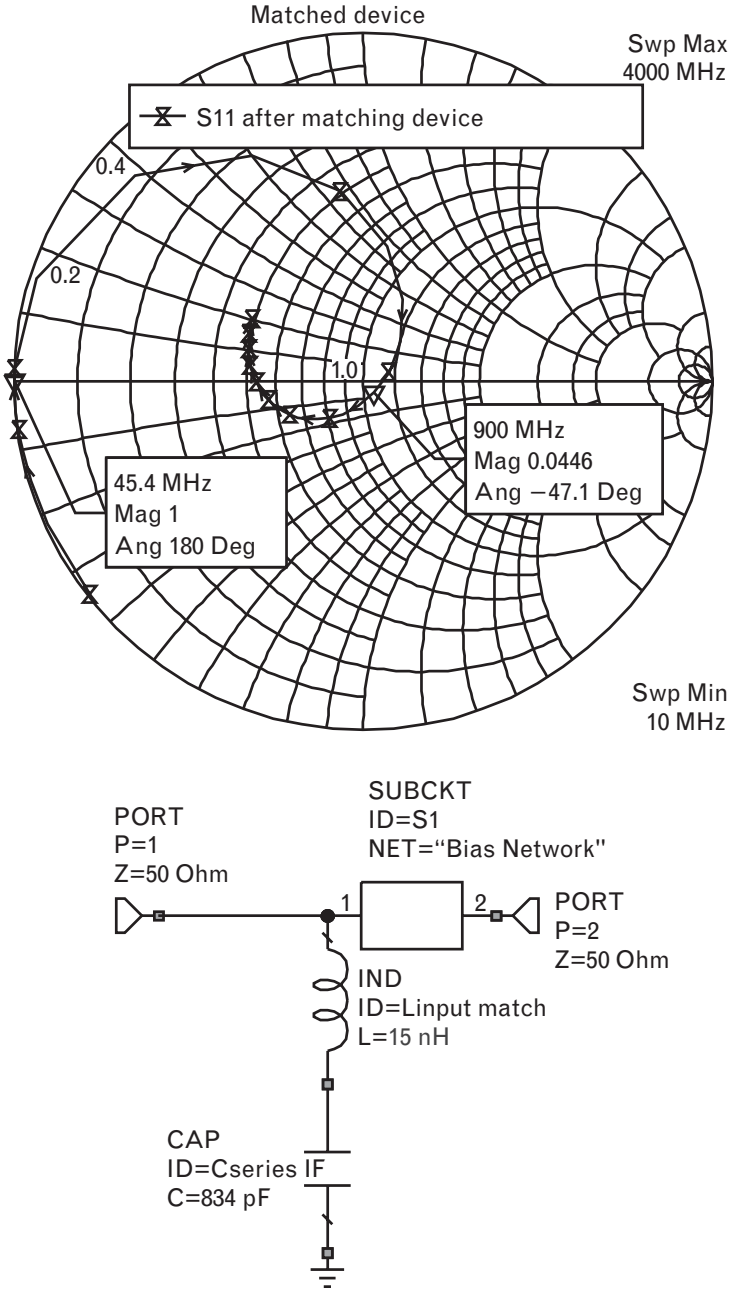


FIGURE 7.27
The small-signal input match to the BFP640 HBT at 900 MHz, and its input matching network.

To be able to apply the estimates for conversion gain and LO power developed in the previous section, we need to understand a little more about the transistor itself. We saw in Chapter 3 that g_0, the average or dc component of g_m, is given by

$$g_0 \approx \frac{qI_E}{kT} = \frac{I_E}{26mV} \tag{7.27}$$

Thus if this transistor is biased at 13-mA average dc current, the average value of transconductance g_m in (7.16) is $g_0 = 13/26 = 0.5$. If we assume the device is nonideal so the transconductance waveform is not square but more half-sinusoidal, then the peak value g_{MAX} is π times the average g_0 (i.e., $\pi/2$).

Now if the collector current is a half-wave rectified sinusoid as for typical class-B operation, then the peak value of the current swing is also π times the average value, or 13π mA (41 mA). Therefore, the necessary zero-to-peak voltage swing at the base to drive the collector current class-B is $\Delta V_{IN} = V_{MAX} - V_{MIN} = I_P/g_0 = 13\star\pi/0.5$ or 82 mV.

We see from Figure 7.26 that at low frequencies, s_{11} lies close to a line of constant conductance with shunt capacitance. At 100 MHz, the normalized input conductance is about 0.42, or 120Ω, in parallel with a normalized susceptance of $j0.21$, or 6.7 pF. Thus, at low frequencies, to a first approximation, the base appears to be a parallel connection of $r_\pi = 120\Omega$, and $C_\pi = 6.7$ pF.

The emitter resistor of the transistor is just $r_E = 1/g_0 = 2\Omega$. The frequency at which the current gain becomes unity is therefore $f_T = 1/(2\pi C_\pi r_E) = 1/(2\pi \bullet 6.7\star 10^{-12}\star 2) = 11.9$ GHz. The low-frequency current gain is then simply $h_{fe0} = r_\pi/r_E = 60$, so from Chapter 3, the 3-dB rolloff frequency can be calculated from the gain bandwidth product as $f_{3\text{–dB}} = f_T/h_{fe0} = 11.9\star 10^9/60 = 198$ MHz. Above this frequency, the input capacitance starts to dominate over r_π, and the base begins to look increasingly capacitive. The series base resistance r_b then becomes more important. This is observed in Figure 7.26, where the input lies increasingly along a circle of constant resistance on the Smith chart as the frequency increases. At 4 GHz, we estimate the normalized resistance to be about 0.5, or 25Ω. The reactance change between 3 and 4 GHz is $+j0.137$ (from $+j0.0462$ to $+j.1832$) or $+j6.85\Omega$. Since at this frequency the series reactance of the base model is simply that of the base inductance $j\omega L_b$ in series with the input capacitance $-j/\omega C_\pi$, and the reactance change due to the capacitance is $+j2.0$, the remainder, $+j4.85$, must be due to the inductance. From this, we calculate the base inductor to be 0.77 nH.

We can now optimize this simple base input circuit so its input impedance matches the device s_{11} across frequency. The optimized equivalent circuit model is shown in Figure 7.28, and the resulting input reflection

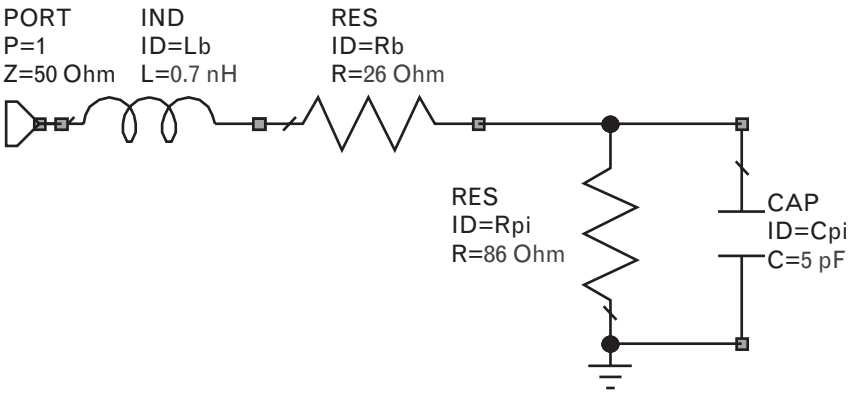


FIGURE 7.28
The small-signal equivalent input circuit model of the BFP640.

coefficient is plotted in Figure 7.26. Although this approach is very crude since the entire SOT343 package model has been collapsed into this model with the device, it serves to obtain a first-order starting point in understanding the device, at least at these moderate frequencies.

These values can now be used in (7.21) to calculate the LO drive power required to achieve such a swing. If we assume an 855/900-MHz operating frequency, we obtain by substituting the values in Figure 7.28

$$P_{LO} = \frac{1}{2} \star 26\left(2\pi \star 855 \star 10^{6} \star 5 \star 10^{-12}\right)^{2} \left(0.082\right)^{2} \tag{7.28}$$

or –12 dBm. Using (7.26) with $k = 0.5$ for a half-wave sinusoidal g_m waveform and a 50-Ω IF output impedance, we obtain for the conversion gain

$$G_T = \frac{1}{4}\left(\frac{0.5\pi/2}{2 \star \pi \star 900 \star 10^{6} \star 5 \star 10^{-12}}\right)^{2} \frac{50}{26} = 371 \tag{7.29}$$

or 25.7 dB.

However, when we begin to drive the device as a mixer we quickly find that the input match is quite different than expected. On investigation, we discover that the reason for this is the nonlinear behavior of the mixer circuit when driven class-B by a relatively large LO signal. The behavior of the HBT input when driven from small signal into compression at 900 MHz is shown in Figure 7.29.

We can see at –20 dBm the same small-signal input reflection coefficient assumed earlier. However, the device input impedance is a strong function of input drive level, and moves towards the open-circuit part of the Smith chart at larger RF drive levels. The shunt inductive input match no longer matches the device when it is driven at +10 dBm, and instead a series inductance is now more appropriate to achieve a reasonable match to input frequencies around 900 MHz. This is rather unfortunate, since the original shunt inductor could be combined with a series capacitor to

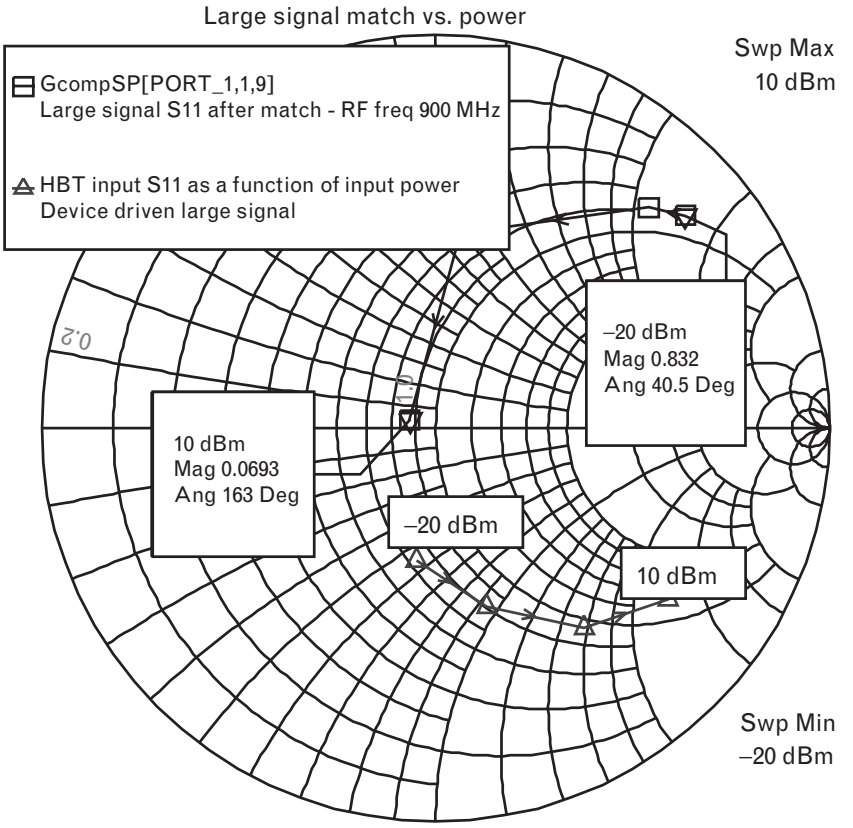


FIGURE 7.29
The simulated large-signal input reflection coefficient of the BFP640 as a function of input drive power, at 900 MHz.

resonate at 45 MHz and short-circuit the IF frequencies at the input. The series inductor cannot achieve such a short circuit at the IF, so in practice, a more complex input circuit would be required than we will use to illustrate the basic principles here.

An inductor of 27 nH in series with the base achieves the RF match shown in the second curve in Figure 7.29. We can see that the optimum match depends on how hard the device is driven large-signal. There is no single best fit, so the RF match should be optimized once the correct LO drive power has been determined.

The final mixer circuit is shown in Figure 7.30. We combine the RF and LO through two series capacitors that serve as a crude but simple coupler. The RF is fed through a large 50-pF capacitor to the base input matching circuit. RF loss is limited because the size of the capacitance provides little reactance. On the other hand, the LO is fed through a much smaller 0.5-pF capacitor. This will reflect most of the available LO power, coupling only a small portion of it into the base so as not to load the RF input. A small capacitor is also required to help isolate the LO and RF signals from each other. This, of course, lowers the LO power that reaches the base, because of the severe mismatch, but we will see that it is still sufficient to drive the mixer. This places a fairly severe constraint on the acceptable

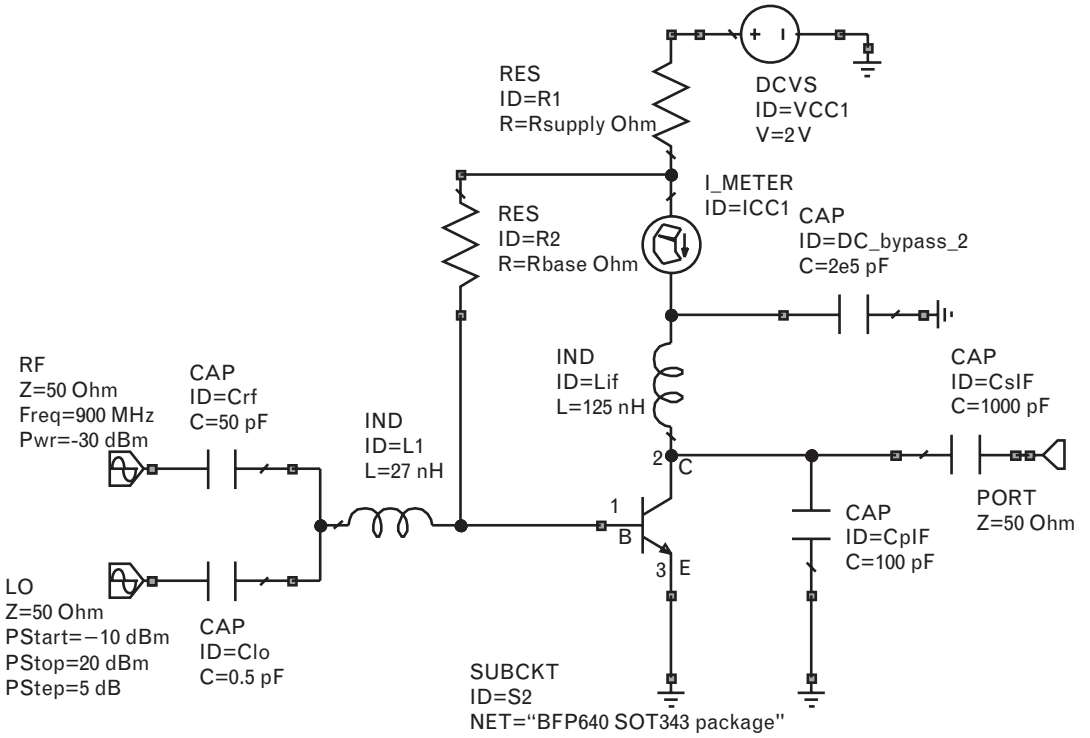


FIGURE 7.30 *The 900-MHz bipolar mixer using the BFP640.*

VSWR to the LO circuit, and requires a much higher LO power level than would be required if the LO were instead matched to the mixer transistor. If this approach proves unacceptable, then a balun such as a transmission line transformer could be used instead to eliminate the reflective loss in the LO path.

At the output, the RF choke of 125 nH and the shunt capacitor at the collector are chosen to resonate at the IF frequency. The choke provides a good feed for the dc collector current, and resonates with the 100-pF capacitance to maximize the IF signal while providing the necessary short circuit to the RF and LO.

We bias the device close to class-B and will use the LO to drive the device large signal. The series collector bias resistor is set to 80Ω and the collector-base resistor to 1,600Ω. The current is plotted in Figure 7.31, which shows how the quiescent collector current increases with LO power. Of course, the large LO powers shown in the figure are unnecessary for mixer operation, as we have calculated above and will shortly verify.

The operation of the mixer is best examined at first in the time domain. The waveforms of the collector current at LO drive levels of −10, 0, and +10 dBm are shown in Figure 7.32. The RF input power is held

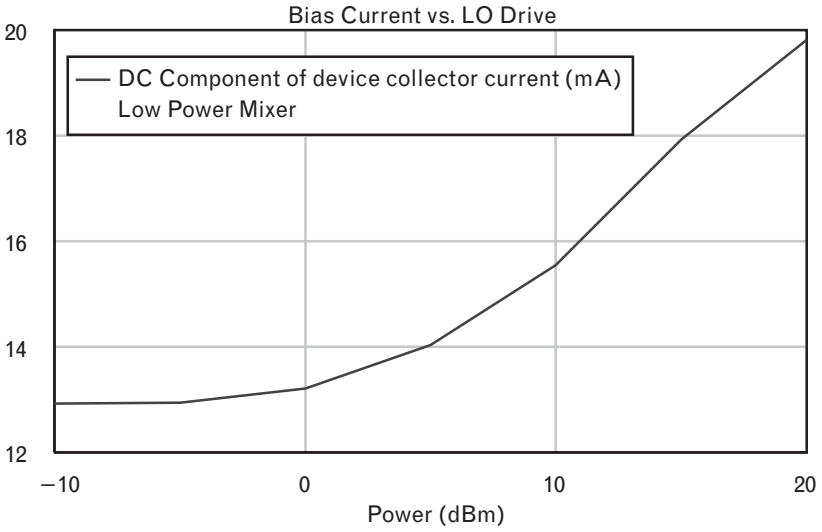


FIGURE 7.31
Bias current of the mixer in Figure 7.30 as a function of LO drive.

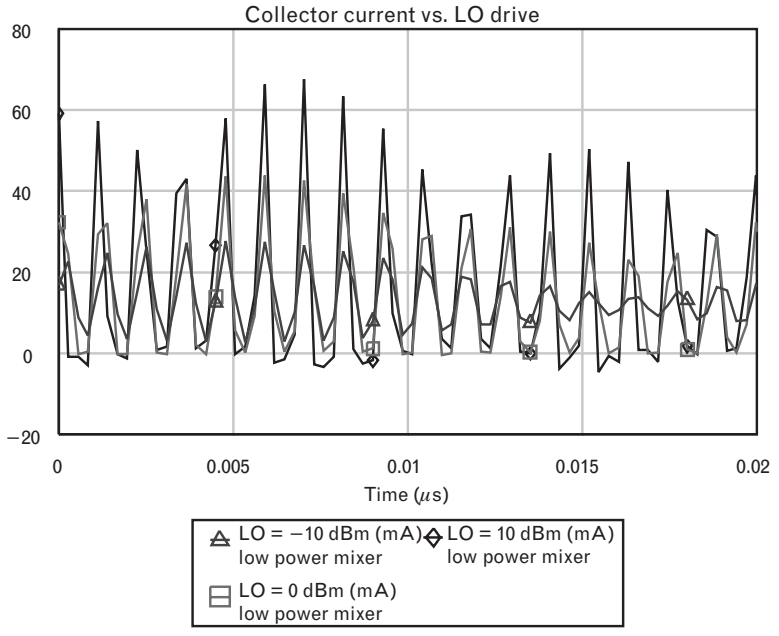


FIGURE 7.32
The collector current of the BFP640 mixer at three different LO drive levels. The RF input power was held constant at −30 dBm.

constant at –30 dBm. It is apparent from these why the average bias current increases as the LO drives the device more large-signal, as we have just noted. At –10 dBm, the envelope of the LO output collector current is modulated fairly uniformly by a slowly varying beat frequency, with period of 0.022 μs, corresponding to the 45-MHz IF. The amplitude of the LO is too small to drive the zero-to-peak current swing larger than the quiescent bias current of 13 mA, so the device is not driven into cutoff on negative half-cycles. However, as the LO input drive increases, the device is driven

increasingly in class-B operation, switching off on the negative peaks of collector current. The amplitude of the envelope also increases together with the IF output power.

Therefore, the conversion gain at first increases, as the amplitude of the output envelope increases with LO drive. Ultimately, however, the device will begin to saturate on positive-going current peaks, and the conversion gain will start to drop as the envelope fluctuation becomes less. Figure 7.33 shows this behavior. The peak conversion gain of almost 24 dB occurs at an LO drive of 10 dBm. Although this gain agrees remarkably well with our calculated value of 25.7 dB, the LO power required is greater by more than 20 dB. However, this is entirely expected, since by using a 0.5-pF coupling capacitor at the base to couple the LO, s_{11} on the LO port has simulated magnitude 0.98. From Volume I, Chapter 2, the LO mismatch loss is then $10\log_{10}(1-|s_{11}|^2)$ or –14 dB, representing the loss in delivered LO power as a result of the capacitance compared to a perfect match. On the other hand, because the RF is matched, the conversion gain agrees much more closely with our calculation.

The remaining simulations are done with an input LO power of –10 dBm, where the conversion gain is about 14.5 dB. Figure 7.34 shows the output spectrum of the mixer with an RF input power of –30 dBm. In addition to the IF output of –15.5 dBm at 45 MHz, the (2,2) term at 90 MHz is also obvious and has an output power of –63.5 dBm, some 48 dB lower. The RF and LO feedthrough, and their third-order mixing products, are also rather high, and could be reduced by better filtering at the output.

Both the conversion gain and the IF output power versus RF power level are shown in Figure 7.35. The conversion gain remains constant until

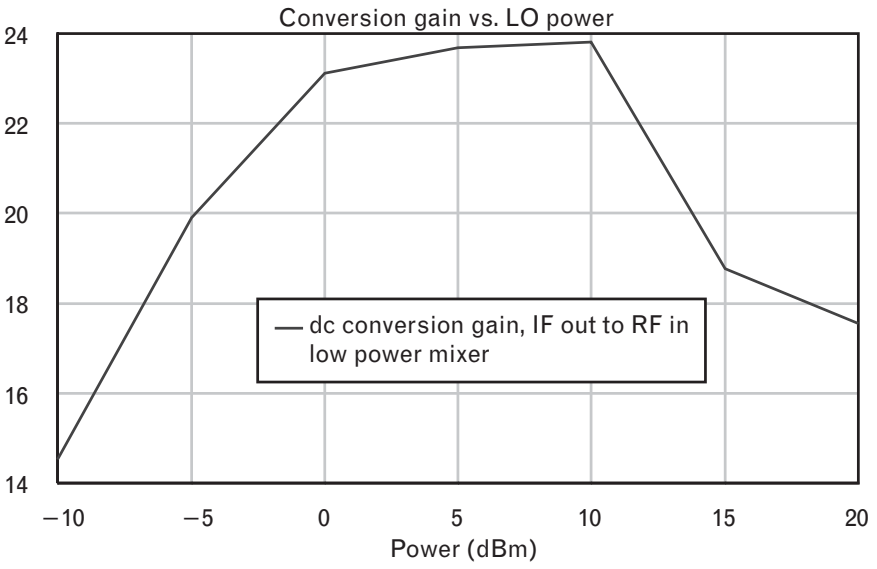


FIGURE 7.33
The mixer conversion gain as a function of LO input power.

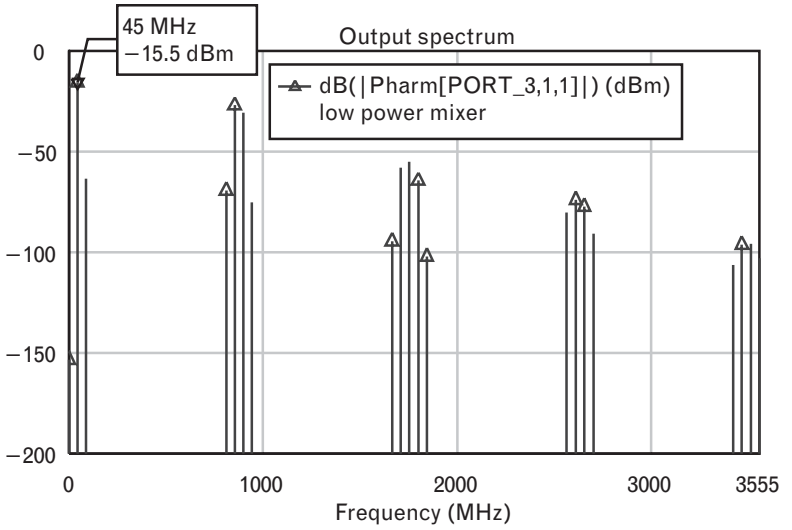


FIGURE 7.34
The output spectrum of the BFP640 mixer. RF input power was −30 dBm and LO power −10 dBm.

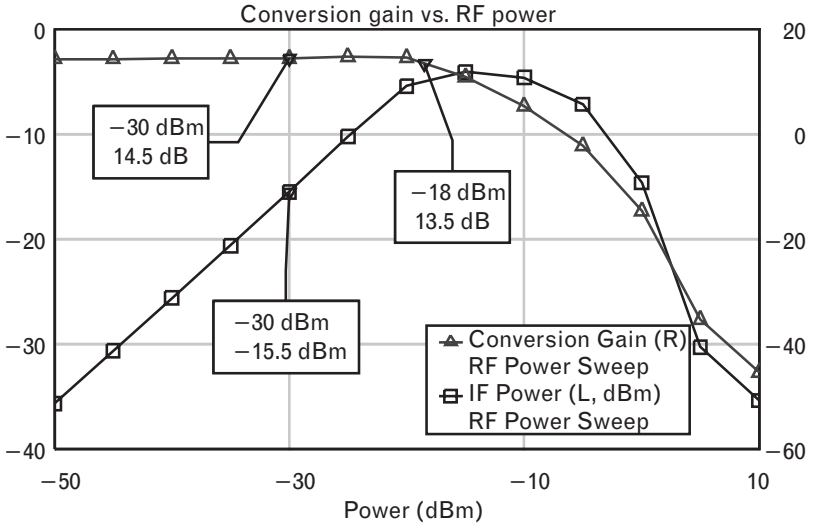


FIGURE 7.35
The conversion gain and IF output power of the BFP640 mixer as a function of RF input power.

the device starts to go into compression at around −18 dBm RF input. The IF output power at the 1-dB compression point is then around −5 dBm. The output 1-dB compression point increases to +3.5 dBm if the LO input power is increased to 5 dBm but starts to decrease again as the LO power is further increased. This compression point is relatively low and indicative of the nonlinear, exponential behavior that is inherent in any bipolar device, especially when used for frequency conversion. The third-order output intercept point could also be derived from simulations by driving the device with two RF tones (together with the LO) and extrapolating the level at which their third-order IF product would be equal in power to the desired IF component.

7.3.1.2 Active FET mixers

The analysis of an FET mixer can now follow a similar procedure to that for the bipolar. In Figure 7.36 the equivalent circuit model for an FET replaces the bipolar transistor of Figure 7.25, where the only change is that the input is now represented by a series R-C circuit, where the input resistor R_{GS} is the sum of the gate, source, and intrinsic FET resistors $R_S + R_i + R_G$ and the capacitance is now the gate-source capacitance C_{GS}. Once again, the input matching network transforms the LO and RF source impedances to the FET input resistance, and the inductance of the matching network effectively resonates out the gate-source capacitance of the FET.

We can derive identical expressions to those of the bipolar transistor for the local oscillator power necessary to drive the FET class-B, and for the conversion gain:

$$P_{LO} = \frac{1}{2}\left(R_S + R_i + R_G\right)\left(\omega_{LO} C_{GS}\right)^2 \left(V_{MAX} - V_{MIN}\right)^2 \tag{7.30}$$

$$G_T = \frac{1}{4}\left(\frac{kg_{MAX}}{\omega_{RF} C_{GS}}\right)^2 \frac{R_L}{\left(R_S + R_i + R_G\right)} \tag{7.31}$$

As before, k varies between 0.5 and $2/\pi$ depending on the shape of the transconductance waveform. The zero-to-peak LO voltage swing necessary to drive the device class B is now from pinch-off to the point of forward conduction, so we typically take $\Delta V_{IN} = V_{MAX} - V_{MIN} = |V_p| + 0.5$ for a MESFET to avoid forward conduction. Driving the gate into forward conduction is not only potentially destructive for the FET, it rapidly increases the noise figure and decreases the conversion gain. Avoid it! Conversely, if LO power is at a premium, a device with a lower pinch-off voltage requires a smaller swing at the gate to achieve the same transconductance variation. HEMTs would be a good choice in this scenario.

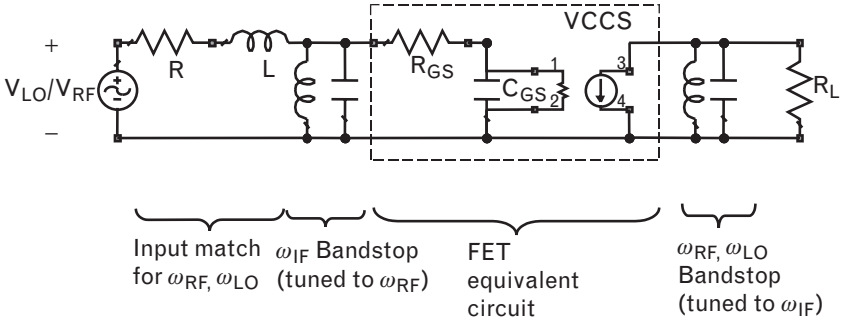


FIGURE 7.36
An active FET mixer with the FET replaced by its small-signal equivalent model.

Active FET mixer example

Consider a mixer with the same functional requirements as the example in Section 7.3.1.1, but using an FET instead of a bipolar transistor. We will use the same linear small-signal FET we previously used for the design of a balanced amplifier, the ATF-54143 pHEMT from Agilent biased at 3V, and assume a nominal class-B generated dc current of 60 mA. Although the device is packaged, at RF frequencies we will initially approximate the package effects by series inductance. As for the bipolar, we can derive an equivalent circuit model for the HEMT input by examining the input *S*-parameter. We can determine that to first order $R_S + R_i + R_G = 12.6\Omega$, $C_{GS} = 4.8$ pF, and $g_m = 450$ mS. This particular HEMT is an enhancement mode device, in which the forward voltage swings from a threshold (i.e., pinch-off) voltage of +0.3V up to about 0.7V to achieve a peak-to-peak current swing of 120 mA. Therefore, a bias of +0.3V and a zero-to-peak voltage swing of just 0.4V will drive this device class-B, between pinch-off and full current. Substituting into (7.30) and (7.31) at 855 MHz and 900 MHz, respectively, we obtain similarly to before

$$P_{LO} = \frac{1}{2} * 12.6 * \left(2\pi * 855 * 10^{6} * 4.8 * 10^{-12}\right)^{2} (0.4)^{2} \tag{7.32}$$

or –1.7 dBm, and

$$G_T = \frac{1}{4}\left(\frac{0.5 * 0.450}{2 * \pi * 900 * 10^{6} * 4.8 * 10^{-12}}\right)^{2} \frac{50}{12.6} = 68 \tag{7.33}$$

or 18 dB. These results are of the same order as those for the active bipolar mixer. Some MESFETs have very small gate capacitances, which can give seemingly high gain results at low frequencies. However, such a gain is somewhat fictitious since it assumes that we can conjugately match such small capacitances at RF frequencies, a feat made simpler at microwave frequencies where their reactance is more reasonable.

If instead we design a mixer using this device at 2.5 GHz, with a local oscillator at 2.4 GHz and an IF at 100 MHz, we obtain instead a LO power of 7.2 dBm and a conversion gain of 9.5 dB. The LO power is increased by 20 dB per decade increase in frequency and the conversion gain drops by 20 dB. This is due to the dependence on $\omega^2 = (2\pi f)^2$ in each equation.

The circuit schematic of such a mixer is shown in Figure 7.37. The LO and RF signals at 2.4 and 2.5 GHz, respectively, are capacitively coupled as before to the gate of the HEMT through a 25-Ω quarter-wave line that matches the device 12-Ω input impedance at this frequency. The necessary short circuit to the IF frequency at 100 MHz is provided by a shunt

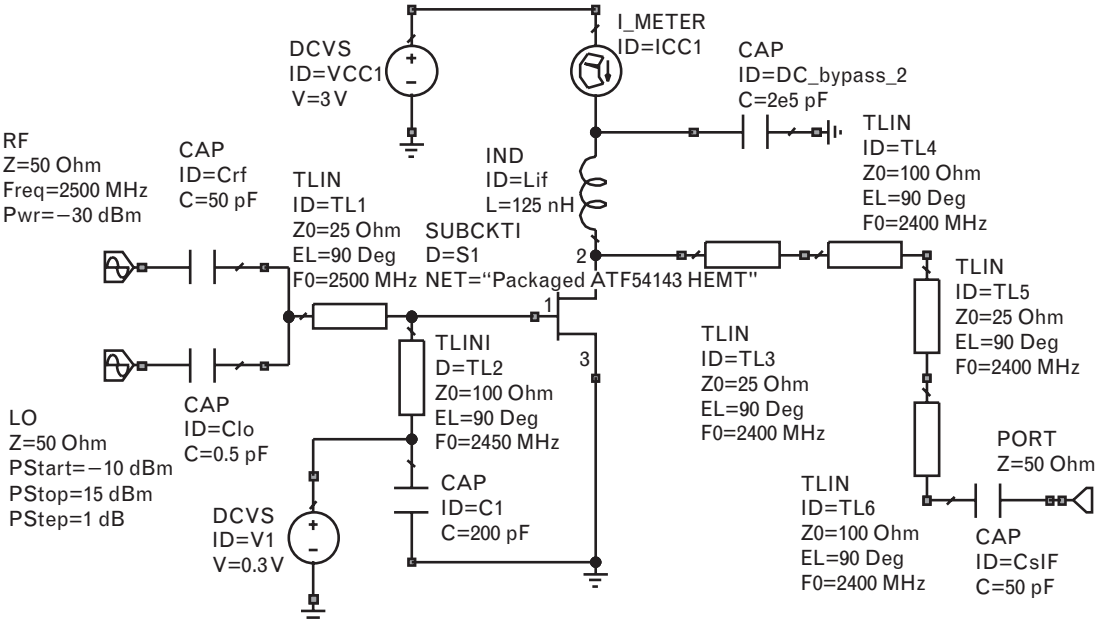


FIGURE 7.37 *Circuit of an active microwave FET mixer using the ATF-54143 HEMT.*

short-circuited high-impedance transmission line, quarter-wave long at 2.5 GHz. The short circuit is achieved by the 200-pF capacitor to ground at the far end of the line. This shorted line, therefore, presents an open circuit at the RF and LO frequencies, and has no impact on the input match to either the RF or LO, but at 100 MHz the line is electrically short in length and appears as a short circuit to the IF. The gate bias can also be fed through this line, brought in on the top of the 200-pF capacitor that can serve as a bypass capacitor. The impedance of the transmission line is chosen quite high so that it appears inductive to the bias network. The dc bias voltage is set at 0.3V (the threshold voltage) so the device operates in class-B mode.

The output circuit needs to provide a path for the IF component of the drain current to the load, while short-circuiting the RF and LO frequencies. Here, this function is provided by an IF filter, which consists of two cascaded 25-Ω/100-Ω line sections. Each line section is one quarter-wave long at 2,400 MHz. At this frequency and working from the load, the first 100-Ω line transforms the 50-Ω load to 200Ω, and the following 25-Ω line transforms this 200Ω down to 200/64Ω, or approximately 3Ω. (The second transformation turns ratio is 25:200 or 1:8, so the impedance transformation is 1:64.) Consequently, the entire output filter of two such sections transforms the 50-Ω load to 3⋆3/50Ω, about 0.2Ω. This provides the necessary short circuit at the drain for the fundamental RF and LO frequencies, while at the IF frequency the electrical length of the entire filter

still has negligible impact. No matching is provided at the IF frequency, since the FET already has conversion gain.

The results of the mixer harmonic-balance simulation are shown in Figures 7.38 to 7.40, respectively. Figure 7.38 shows the conversion gain of the mixer as a function of LO power. The peak conversion gain of around 17 dB occurs with +14 dBm LO power at the gate capacitor input (much less at the gate itself due to the mismatch loss of this capacitor). The circuit and IF filter provide good LO rejection at the output since the simulated LO output level is only –32 dBm.

The drain current waveforms of Figure 7.39 illustrate very clearly the class-B behavior of the mixer. The RF input level was held constant at –30 dBm for this simulation. As the LO drive increases, the beat-frequency envelope of the IF waveform may be clearly seen, superimposed on the RF. The IF frequency is 100 MHz, corresponding to a period of 0.01 μs.

The simulation of the RF response was done with an input LO power of +10 dBm, where the conversion gain was close to its peak. Figure 7.40 shows the conversion gain as the input RF signal is increased. The input 1-dB compression point of this mixer is at –3.8 dBm, corresponding to an output 1-dB compressed IF power of 11.2 dBm.

This is an excellent result that turns out to be highly dependent on the quality of the output LO short circuit, which allows a very large output current swing. Although this translates into high dc power consumption as well, the benefits of properly terminating the LO leakage are higher conversion gain and higher intercept point. Using a less effective output termination rapidly reduces the mixer 1-dB compression point.

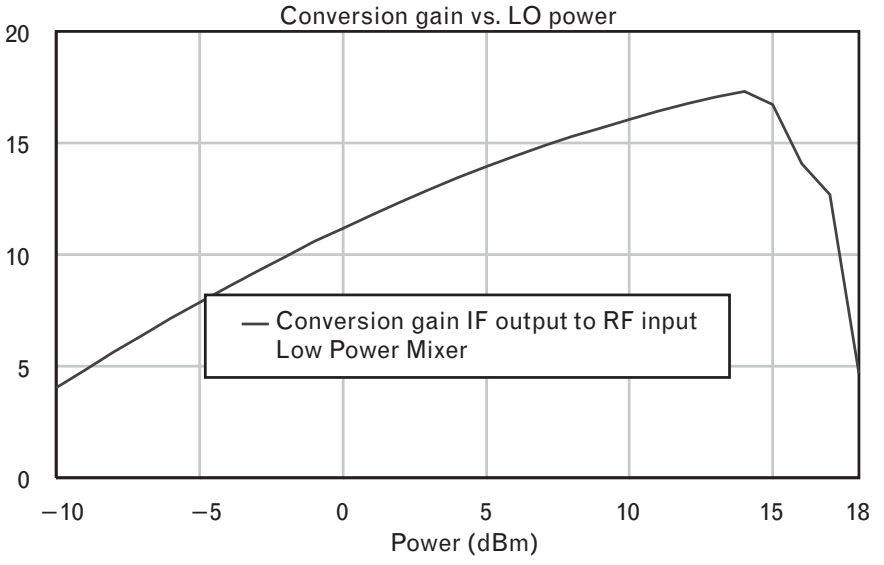


FIGURE 7.38
The mixer conversion gain as a function of LO input power.

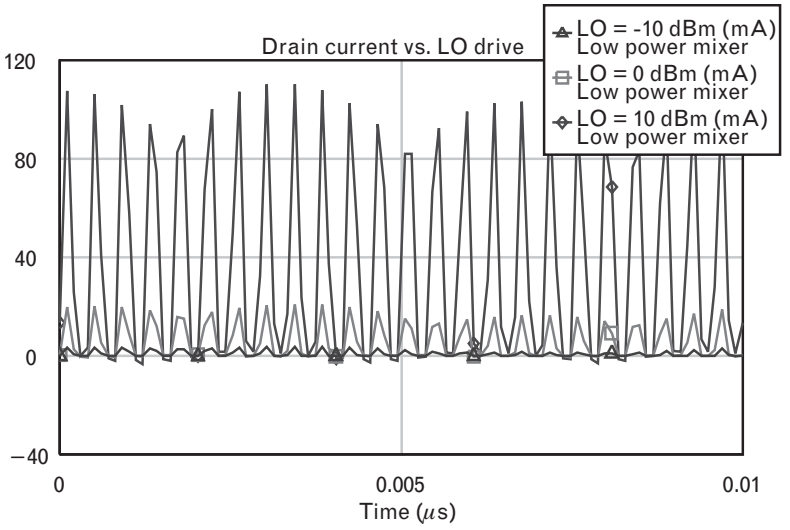


FIGURE 7.39
The drain current of the ATF-54143 HEMT mixer at three different LO drive levels. The RF input power remained constant at −30 dBm.

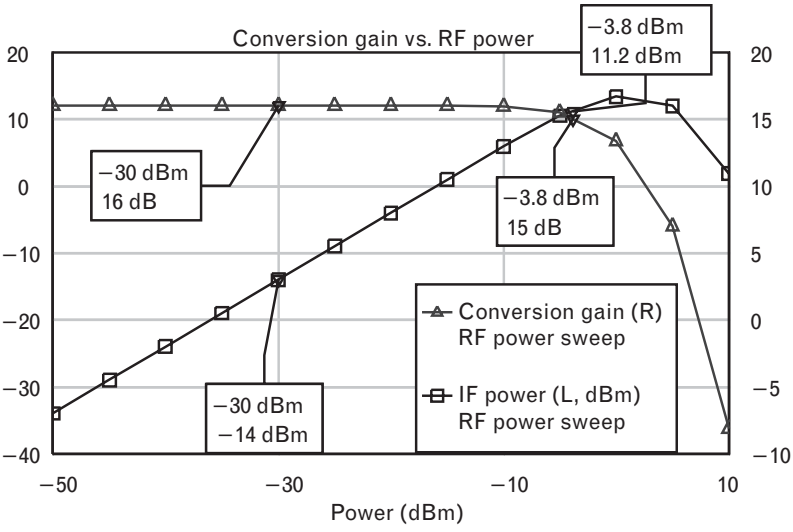


FIGURE 7.40
The conversion gain and IF output power of the ATF-54143 HEMT mixer as a function of RF input power.

7.3.1.3 The Gilbert cell mixer

We have seen earlier that a bipolar transistor can be used in single- ended form and achieve mixing through its transconductance nonlinearity. However, it can be difficult to achieve the necessary isolation between ports, particularly if the IF is low and is easily amplified by the transistor itself.

The benefits of a double-balanced mixer, with full integration on a single chip, can be achieved with the Gilbert cell mixer. Originally intended by its designer in 1968 to be used as a four-quadrant analog multiplier [5], the Gilbert cell mixer can be used in a switching mode for mixing. Using

differential pairs of bipolar transistors without the need for external baluns, the mixer offers high gain, wide bandwidth, and low power consumption—so has become a popular choice for integrated mixers. The main drawback is that because of the exponential nature of the bipolar transfer relationship, the intermodulation performance can be bettered by other mixers such as the resistive FET mixer, which we will cover later.

The general schematic of a PMB2335 Gilbert cell mixer from Infineon technologies is shown in Figure 7.41. Although the Gilbert cell has a number of variants in terms of types of input feed (single ended, differential, low or higher impedance, and so on), the principles are always the same. In this implemenation the RF voltage is fed differentially to the emitters of a differential pair Q5/Q6, whose bases are held at a constant bias voltage.

This pair sources a total constant emitter current I_{ee}, while the RF signal modulates each collector current individually. Due to the relatively low supply voltage of just a few volts, the normal current source provided by current sinking transistors between the emitter nodes and ground has been omitted. The advantage of using a current source there would be that the RF drive could be from an unbalanced source, since the current source prevents any even mode current flowing in the two RF transistors, and the Q5/Q6 emitter voltage could simply float up and down independently of the current.

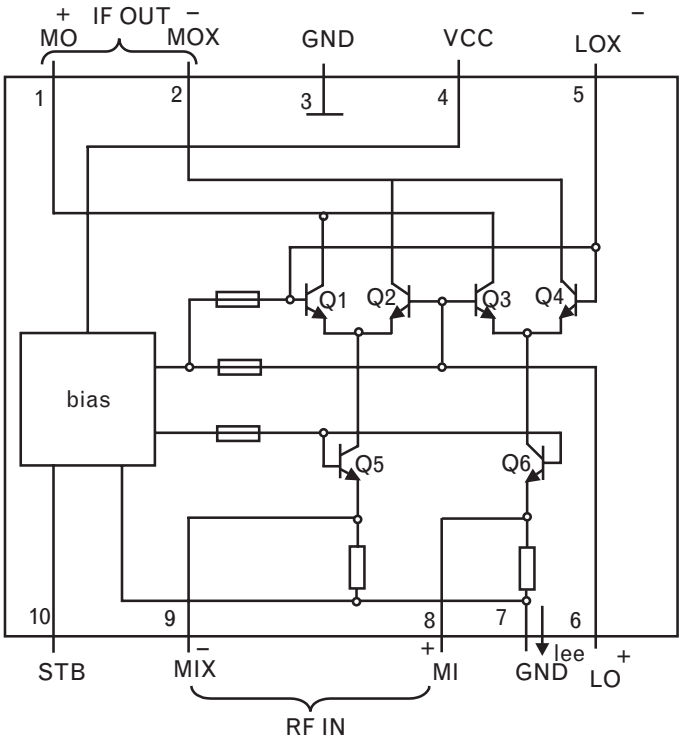


FIGURE 7.41
Circuit schematic of a Gilbert cell mixer. (Courtesy Infineon Technologies.)

Series resistors are sometimes used with Q5 and Q6. These emitter degeneration resistors provide feedback for the lower differential pair and improve the linearity of the mixer. They also allow larger RF signal swings at the input, increasing the dynamic range and increasing the mixer intercept point, although they will degrade the noise figure. Either way, the emitters of Q5/Q6 form a virtual ground for the RF.

The collectors of Q5/Q6 form the emitter load of two pairs of cross-coupled differential amplifiers (Q1-Q4), whose bases are driven between cutoff and saturation by the differential LO signal. The emitters of these devices form a virtual ground to the LO ensuring there is no LO voltage on the RF devices. The LO switches the collector current between Q5 and Q6 at the LO rate. The output IF current is formed from the difference in the collector currents of the two cross-coupled differential pairs.

Mathematically, if we abbreviate kT/q by V_t, then we can write expressions for each transistor of the form

$$I_C(Q5) \approx I_S \exp\left(V_{be}(Q5)/V_t\right) \tag{7.34}$$

so that

$$V_{be}(Q5) = V_t \log\left(I_C(Q5) / I_S\right) \tag{7.35}$$

Such an expression ignores RF effects such as the output capacitance of the devices, but it is helpful in providing a first-order understanding of device operation. Now since the RF voltage is applied differentially to the emitters of Q5 and Q6, we have

$$\begin{aligned} V_{RF} &= V_{eb}(Q6) - V_{eb}(Q5) \\ &= V_t \log\left(I_C(Q5)/I_C(Q6)\right) \end{aligned} \tag{7.36}$$

or

$$I_C(Q5)/I_C(Q6) = \exp\left(V_{RF}/V_t\right) \tag{7.37}$$

Now using $I_{ee} = I_C(Q5) + I_C(Q6)$ with (7.37), we obtain

$$I_C(Q5) = \frac{I_{ee}}{1+\exp\left(-V_{RF}/V_t\right)} \tag{7.38}$$

and

$$I_C(Q6) = \frac{I_{ee}}{1+\exp\left(+V_{RF}/V_t\right)} \tag{7.39}$$

However, the differential output current is formed from the collector currents of the output differential pair and is given by

$$\Delta I_{OUT} = I_C(Q1) + I_C(Q3) - I_C(Q2) - I_C(Q4) \tag{7.40}$$

Using expressions of the form of (7.39) for each of these components, we obtain

$$\begin{aligned}
\Delta I_{OUT} &= \frac{I_C(Q5)}{1+\exp(V_{LO}/V_t)} + \frac{I_C(Q6)}{1+\exp(-V_{LO}/V_t)} \\
&\quad - \frac{I_C(Q5)}{1+\exp(-V_{LO}/V_t)} - \frac{I_C(Q6)}{1+\exp(V_{LO}/V_t)} \\
&= I_C(Q5)\left[\frac{1}{1+\exp(V_{LO}/V_t)} - \frac{1}{1+\exp(-V_{LO}/V_t)}\right] \\
&\quad - I_C(Q6)\left[\frac{1}{1+\exp(V_{LO}/V_t)} - \frac{1}{1+\exp(-V_{LO}/V_t)}\right] \\
&= \frac{I_{ee}}{1+\exp(-V_{RF}/V_t)}\left[\frac{1}{1+\exp(V_{LO}/V_t)} - \frac{1}{1+\exp(-V_{LO}/V_t)}\right] \\
&\quad - \frac{I_{ee}}{1+\exp(-V_{RF}/V_t)}\left[\frac{1}{1+\exp(V_{LO}/V_t)} - \frac{1}{1+\exp(-V_{LO}/V_t)}\right]
\end{aligned} \tag{7.41}$$

This expression can be simplified (fortunately!) when we note that

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \tag{7.42}$$

and that the numerator and denominator of each term inside the square brackets above can be multiplied by $\exp(\pm V_{LO}/2V_t)$ so that the terms simplify to

$$\begin{aligned}
\Delta I_{OUT} &= -\frac{I_{ee}\exp(V_{RF}/2V_t)}{\exp(V_{RF}/2V_t)+\exp(-V_{RF}/2V_t)}\left[\tanh(V_{LO}/2V_t)\right] \\
&\quad + \frac{I_{ee}\exp(-V_{RF}/2V_t)}{\exp(-V_{RF}/2V_t)+\exp(V_{RF}/2V_t)}\left[\tanh(V_{LO}/2V_t)\right]
\end{aligned} \tag{7.43}$$

which is just

$$\Delta I_{OUT} = -I_{ee}\left[\tanh\left(V_{RF}/2V_t\right)\right]\left[\tanh\left(V_{LO}/2V_t\right)\right] \tag{7.44}$$

For small values of the local oscillator or RF relative to V_t, the hyperbolic tangent is linear and the output current is simply the product of the LO and the RF voltage. The cell would then serve the original intention of an analog multiplier. The assumption of linearity, of course, is not a particularly good one and indicates one of the key drawbacks of this mixer.

In passing, we should note that this multiplication principle has been used in variable gain amplifiers, whereby the RF signal is applied to the top pairs of transistors and the gain is adjusted by controlling the lower pair. In this case, the control signal is the AGC control voltage that effectively sets the bias current between each pair [6].

In mixer use, the local oscillator signal is driven much larger than V_t, and the hyperbolic tangent term then switches between +1 and –1. As a result, the output current is proportional to a square-wave switching waveform modulated by an RF voltage; that is, from (7.44)

$$\Delta I_{OUT} \approx -I_{ee}\left(V_{RF}/2V_t\right)g_{LO}(t) \tag{7.45}$$

which, of course, is identical to the previous expression (7.3) we have derived for mixer operation. Both the LO-IF and RF-IF feedthrough are very low because of the differential, double-balanced structure, and this is confirmed by the disappearance of the $\cos(\omega_{LO}t)$ and $\cos(\omega_{RF}t)$ terms when the trigonometric product is expanded in this expression. Output filtering at the RF and LO is not necessary, which helps to ensure broadband operation.

The PMB2335 is operable as a double-balanced mixer to 3 GHz. Beyond that frequency, the noise figure degrades rapidly because as we approach the device f_T, the upper pairs lose their ideal switching characteristics. Due to the differential nature of the device, the even spurious terms are kept low and isolation between the RF, LO, and IF ports is excellent. The LO drive requirement is typically –5 dBm, comparable to that for the single-ended active bipolar mixer considered earlier.

At 900 MHz, typical conversion gain is 4 dB, the input IP3 is +3.5 dBm, and the noise figure is 7.5 dB. The noise figure of Gilbert cell mixers is inherently quite high because the lower differential pair provides fairly broadband amplification prior to switching by the upper differential pairs, so that any image noise is aliased to the IF by the switching, and adds about 3 dB on top of the switching losses. This image noise, plus the device thermal and shot noise, are visible at the output during the time that the upper pairs are both instantaneously on during the LO switching transitions, since these pairs then look like a differential amplifier. IF filtering should be used at the RF input to prevent amplification of IF noise, while increasing the LO drive to narrow the transition times also reduces the noise.

This particular mixer features RF drive to the emitters of the lower differential pair, which are connected common-base, rather than common-emitter as in some implementations. The input must then be driven symmetrically, and neither terminal can be RF grounded. However, compared with driving the base, this provides a lower RF input impedance that is better matched to 50Ω, typically important to avoid detuning the off-chip filters that might precede the mixer. Keeping the voltage swing low with a low resistance load can also minimize any intermodulation distortion that might be generated in the filter's tuning diodes. Driving the emitter rather than the base also extends the frequency range since the input Miller capacitance is kept small. The device is packaged in an 8-pin surface-mount package, so it is well adapted to mobile handheld systems.

In CMOS technology, FETs can also be used in Gilbert cell topologies. However, the FETs are not easy to match to each other in CMOS. The threshold voltage mismatch is not important when the LO is large enough to swamp any difference, and mismatch in g_m can be overcome by using source resistors in series with each FET to dominate the equivalent source resistor.

The Gilbert cell multiplier can also be used as balanced modulator. The output is a double-sideband suppressed-carrier modulation that can be used for both AM and FM. Because of the virtual ground between each differential pair, the output signal contains neither the fundamental of the LO nor the RF.

7.3.2 Resistive FET mixers

The transistor mixers we have considered throughout this chapter have all had applied bias, so that when driven into the on-region of the device by the LO signal, an active current flows into the device. This offers the possibility of conversion gain compared with diode mixers.

It is also possible to use an FET in a similar way to a diode mixer, that is, to use the conductance between the drain and source as a mixing element, in much the same way that the diode conductance was used. However, the FET has an advantage over the diode, and that is that it has a third and separate terminal, the gate, at which the LO can be applied.

Diode mixers, because of the exponential relationship between their current and voltage, are strongly nonlinear, with mediocre intermodulation performance. Although the second and third-order intercept points generally increase with applied local-oscillator power, as the switching waveform becomes more square, excessive LO power ultimately increases the noise and conversion loss of a diode mixer.

A resistive FET mixer on the other hand uses the channel resistance, between the drain and source, as a time-varying conductance. With no applied bias voltage at the drain, the LO is applied to the gate to switch the conductance between an on-state and an off-state. Good intermodulation performance results since this conductance varies more linearly with drive

voltage than for a diode. This is important in receivers, since the mixer often handles the largest signal levels of any component, and its response can consequently limit the overall dynamic range. In fact, as long as g_1 in (7.3) is constant with drive level, then perfectly linear mixing results and high input powers can be handled. Because the resistive FET mixer is capable of high distortion-free output power for moderate LO levels, it has become a popular type of mixer.

The principle is shown in Figure 7.42(a). The RF voltage can be applied to either the drain or the source, and the IF filtered from the drain

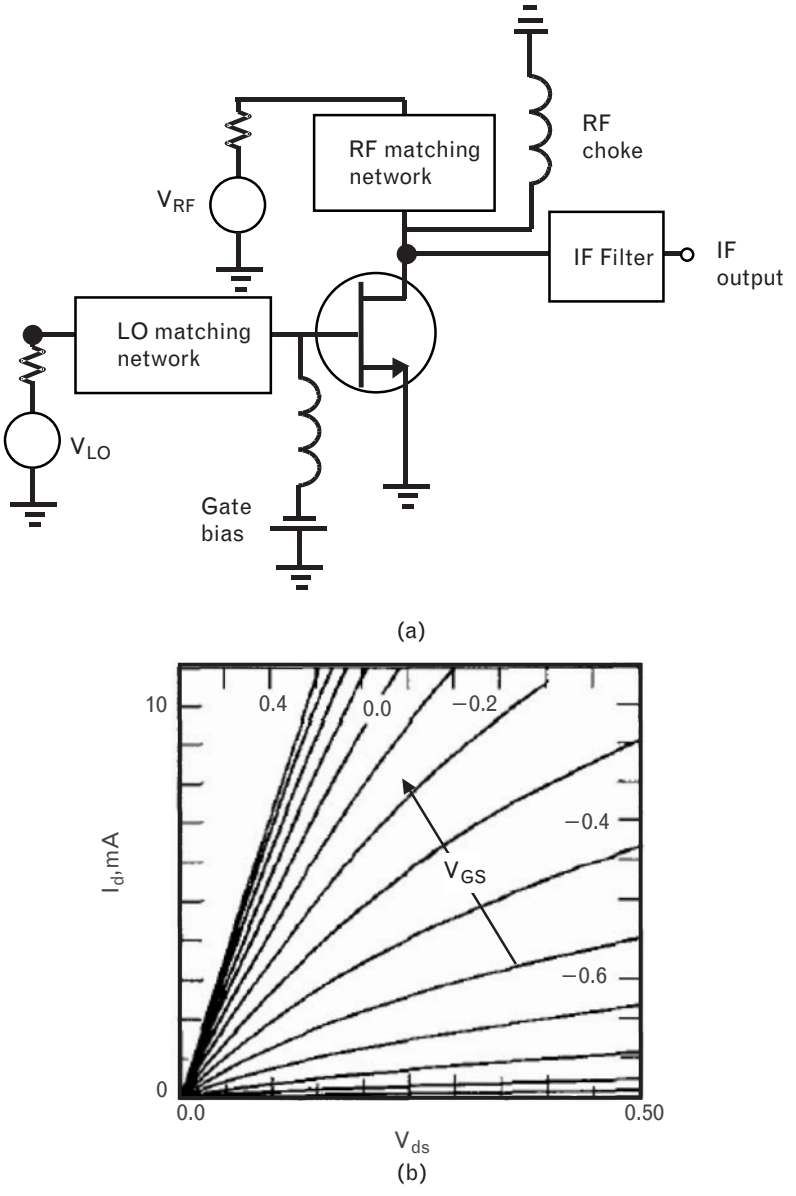


FIGURE 7.42
Principles of a resistive FET mixer: (a) basic circuit, and (b) variation of the channel conductance with applied gate voltage. (From: [7]. © 1987 IEEE. Used with permission.)

current. The I-V curves of the FET in its linear region, below the knee of the curve, are shown in Figure 7.42(b). Because the device has no drain-source bias voltage, the drain-source conductance seen by the RF voltage is simply the slope of the I-V curves around the origin. This slope is switched by the gate at the LO rate. The conductance can swing between zero (when the device is at pinch-off) and several ohms (at the point of forward turn-on). The variation of conductance between these extremes of LO voltage swing is only weakly nonlinear.

Figure 7.43 shows the equivalent circuit of the FET without an applied drain-source voltage. The circuit must hold the dc value of the drain voltage V_d at zero volts with a dc short circuit (the RF choke in Figure 7.42). Without applied drain bias, there is no drain-source current source; the channel is represented by a conductance, which is a function of the gate voltage V_g. The gate-source and gate-drain capacitance are the other dominant elements of the FET. Matching circuits are needed at the gate and drain terminals, respectively, to match the LO and RF, and to short-circuit the other frequency, because when the device has no applied bias, the gate-drain (or gate-source) capacitance is quite large. Since devices with higher f_T have smaller capacitance, the feedthrough can be reduced by using higher frequency devices.

At the gate [Figure 7.43(b)], the LO sees the source and drain terminals as symmetrical if the drain is terminated in a short circuit at the LO frequency and if the source and drain parasitics are approximately equal. The LO short circuit at the drain is required to prevent the drain voltage traversing the knee of the I-V curve and increasing the intermodulation

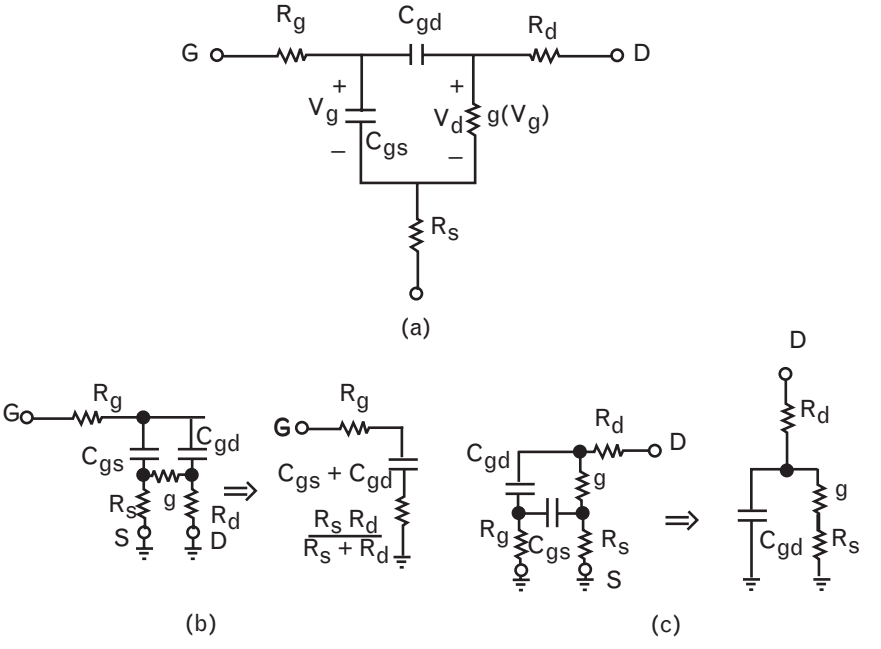


FIGURE 7.43
(a) Equivalent circuit of the FET in the resistive FET mixer: (b) seen from the gate, and (c) seen from the drain. (From: [7]. © 1987 IEEE. Used with permission.)

products. If the RF is close to the LO in frequency, it can be difficult to achieve a good short circuit at the drain, and the balanced structures discussed below should be considered to improve RF-LO isolation. The equivalent circuit at the gate is then modeled by a series R-C circuit in which the channel conductance is not part of the circuit because of the symmetry between the drain and the source noted above.

Matching to the LO is needed at the gate to ensure a full signal swing across it. As with most MESFETs, the gate is not particularly easy to match because of its small series capacitance, but particularly so when the device is in its off state. Sometimes a 50-Ω shunt gate resistor is recommended as a matching element, at the cost of halving the available voltage swing and losing 6 dB in LO power.

In practice, the gate will be biased close to, or even below pinch-off, as this helps to minimize the conversion loss. As in the active single-gate FET mixer, the LO drive level will ideally drive the gate from pinch-off to the point where the gate-source diode just begins to conduct (i.e., the point at which a small component of dc gate current is observed). As long as we avoid forward conduction and reverse breakdown, then driving the gate between two distinct states as fast as possible keeps the conductance variation more linear than for a diode, and the intermodulation response is improved. This is because we then create maximum transconductance variation, maximize the fundamental component g_1, and keep it constant.

The impedance seen at the gate by the LO, and at the drain by the IF, affects both the conversion loss and the spurious response [8]. The conversion loss decreases linearly with increasing LO drive, and will be quite large until the LO signal swing fully drives the gate between pinch-off and the point of forward turn on. Beyond a certain LO drive level, there is no further decrease in conversion loss with LO power. Once the forward turn-on voltage at the gate is reached at the peak of the LO signal swing, the channel conductance cannot increase further beyond some maximum value, and will only add new components to the distortion.

These effects can be seen for a typical resistive FET mixer in Figure 7.44. Clearly, there will be an optimum gate-bias point for each LO power level, with the optimum occurring when the peak gate voltage is just below the transistor turn-on voltage. This bias voltage becomes more negative as the LO power is increased

The impedance seen at the drain by the RF is the time-varying drain conductance in shunt with the drain-gate capacitance, as in Figure 7.43(c). This is similar to a diode equivalent circuit, except that the conductance variation is now more linear. An IF output current is formed because of the relationship in (7.3) between the LO-varying conductance and the RF voltage across it. The RF and IF signals are split by bandpass filters in the drain. The IF load impedance is important as it affects both the conversion loss and the intermodulation distortion products. This implies that in a

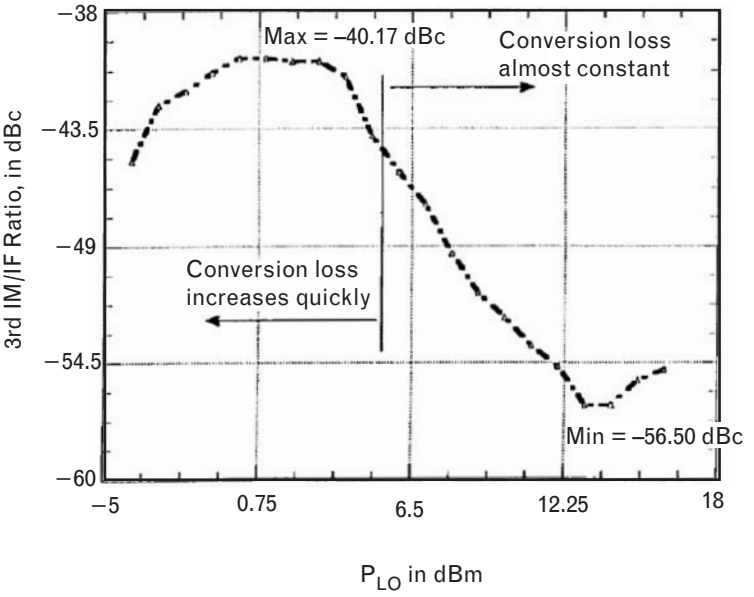


FIGURE 7.44
The variation of the third-order distortion products from the resistive FET mixer (relative to the IF fundamental) as a function of applied LO power. (From: [8]. © 1998 IEEE. Used with permission.)

radio receiver, the filter following the mixer needs to be carefully chosen. As in all mixers, the IF should be short-circuited at the gate.

Resistive FET mixers (like their single-ended active counterparts) can be deployed in balanced configurations. As considered earlier, balanced configurations reject AM noise from the local oscillator and spurious responses involving even LO or RF harmonics. The baluns also provide isolation between the LO, RF, and IF ports—particularly important because of the large gate-drain capacitance in a resistive FET mixer. However, unlike diode mixers where the diode could be geometrically "flipped" to subtract the out-of-phase IF currents at the junction between the single-ended mixers, resistive FET mixers are symmetrical with respect to the drain and source. The subtraction of IF currents now requires a balun at the IF output port as well.

Two configurations to model a single-balanced mixer are shown in Figure 7.45. The input circuits have the same topology as for their diode counterparts, but the output now requires a 180° hybrid in order to subtract the two IF currents in each leg of the mixer. In resistive mixers connected as in Figure 7.45(a) with an input 180° balun at the LO and an output 180° balun at the IF, the in-phase (Σ) function of the balun shown at RF can be achieved by simply tying the drains of each FET together (at the RF frequency) using two capacitors. This allows injection of the RF signal in phase through the capacitive center-tap, to the drains of both devices in parallel. Because the drain impedance is then divided by two, RF matching at the drain may also become unnecessary. Using capacitive coupling in this way also avoids the use of a separate RF filter and helps maintain a high bandwidth. It also provides the necessary output short circuit for the LO since the capacitive center tap is a virtual ground to the LO,

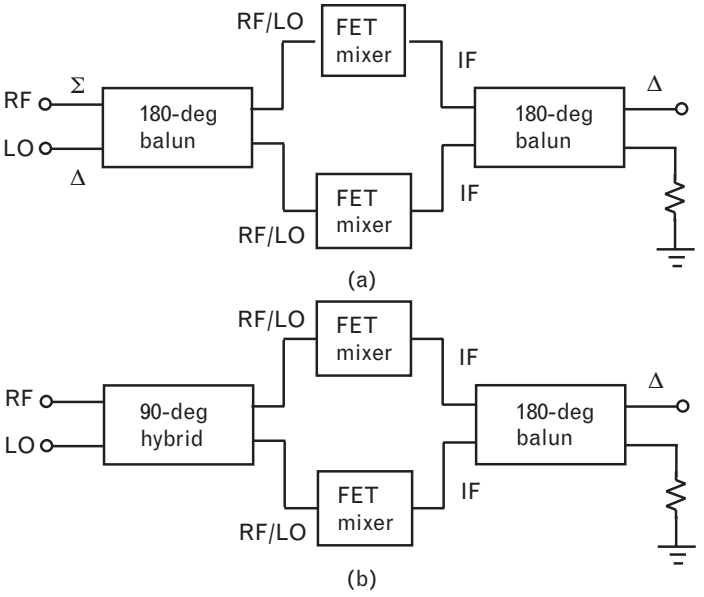


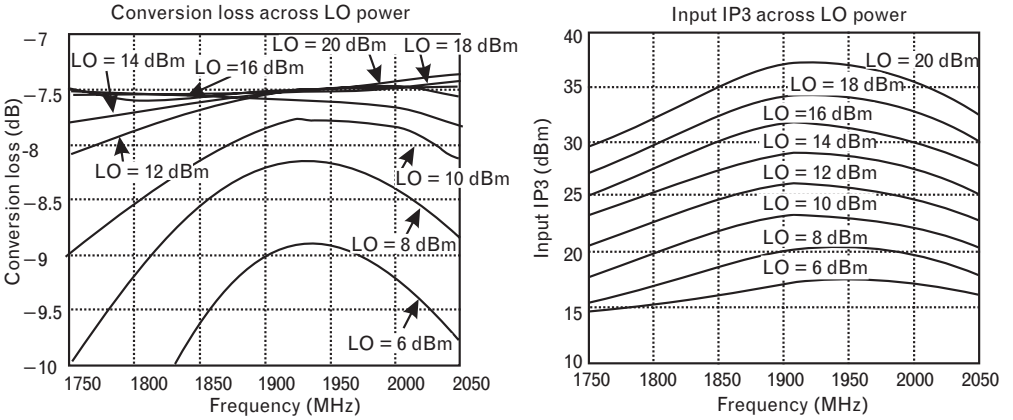
FIGURE 7.45
Balanced FET mixers (a) using 180° baluns and (b) using a quadrature coupler as an input balun.

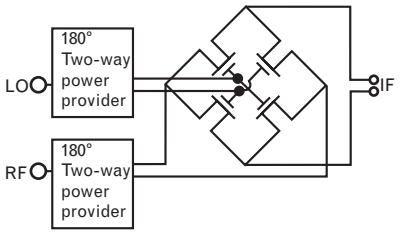
which is still applied out of phase at each gate. LO to RF isolation is thus good. The IF currents are then extracted through an IF filter connected to each drain separately and subtracted in an output transformer (balun).

Because the IF currents are no longer added at a common output node as for diodes, but subtracted in a balun, the spurious responses are opposite those of the diode single-balanced mixer. Thus, in Figure 7.45(a), where the LO is applied out of phase, the (1,2) response is now rejected and no longer the (2,1).

Like diodes, FETs can be packaged as quads for use in double-balanced configurations. The PE4134 quad mixer from Peregrine Semiconductor is one example, and its characteristics are shown in Figure 7.46 for an RF bandwidth extending up to 2 GHz. A typical IF frequency would be 260 MHz. This particular example is fabricated using silicon CMOS technology. Although the Gilbert cell mixer can also be fabricated in CMOS, this mixer consumes no dc power. The conversion loss is around 7.7 dB at 2 GHz with an applied LO drive of +10 dBm. External baluns can be used to drive the package with either single-ended or differential LO and RF signals. In this implementation, the RF signal is fed to the source of the FETs and the IF current is taken from the drain. In monolithic implementations such as this, the partitioning of the IF and RF signals to the source and drain is much easier to achieve. In theory the drain and source are interchangeable, although it would be preferable to take the lower frequency signal of either the IF or RF from the source which, now ungrounded, might as a result have higher parasitic capacitance.

The input third-order intercept point for this double-balanced resistive FET mixer is 23 dBm, an improvement of at least 7 dB over a good double-balanced diode mixer with the same LO power of +10 dBm. The





Spurious response				
	mRF + nLO			
	nLO			
mRF	1	2	3	4
1	5	37	30	50
2	49	55	65	59
3	75	>85	78	80
4	>85	84	>85	>85

Spurious response				
	mRF + nLO			
	nLO			
mRF	1	2	3	4
1	0	32	22	43
2	57	49	61	62
3	64	71	64	65
4	>85	>85	>85	81

Normalized to dB below P_{IF}. LO 217 dBm

FIGURE 7.46 *The PE4134 quad FET mixer and typical performance curves. (Courtesy Peregrine Semiconductor.)*

input intercept point rises to over 35 dBm with +20-dBm LO power, making it ideal for radio base station applications.

7.3.3 Dual-gate FET mixers

We saw in Chapter 5 the use of a dual-gate FET as an amplifier that can be used to compensate for phase distortion over some range of input powers. The device was modeled as a (cascode) connection of two FETs in series, the first in common-source and the second in the common gate.

Such a device, with its four terminals, can also be used as a mixer. In fact, the intercept point of a dual-gate FET mixer is often a few decibels higher than for an active single-gate FET mixer, possibly because distortion is also lower in the mixer for similar reasons as in the predistorter studied earlier.

So far, for both the resistive and active FET mixer, the local oscillator signal has driven the gate of the device to *indirectly* modulate either the conductance of the channel or the transistor transconductance using the device transfer characteristic. Figure 7.47 shows a slightly different mode of operation, in which the local oscillator is injected into the collector or drain of the device to *directly* modulate its output conductance. In order to achieve

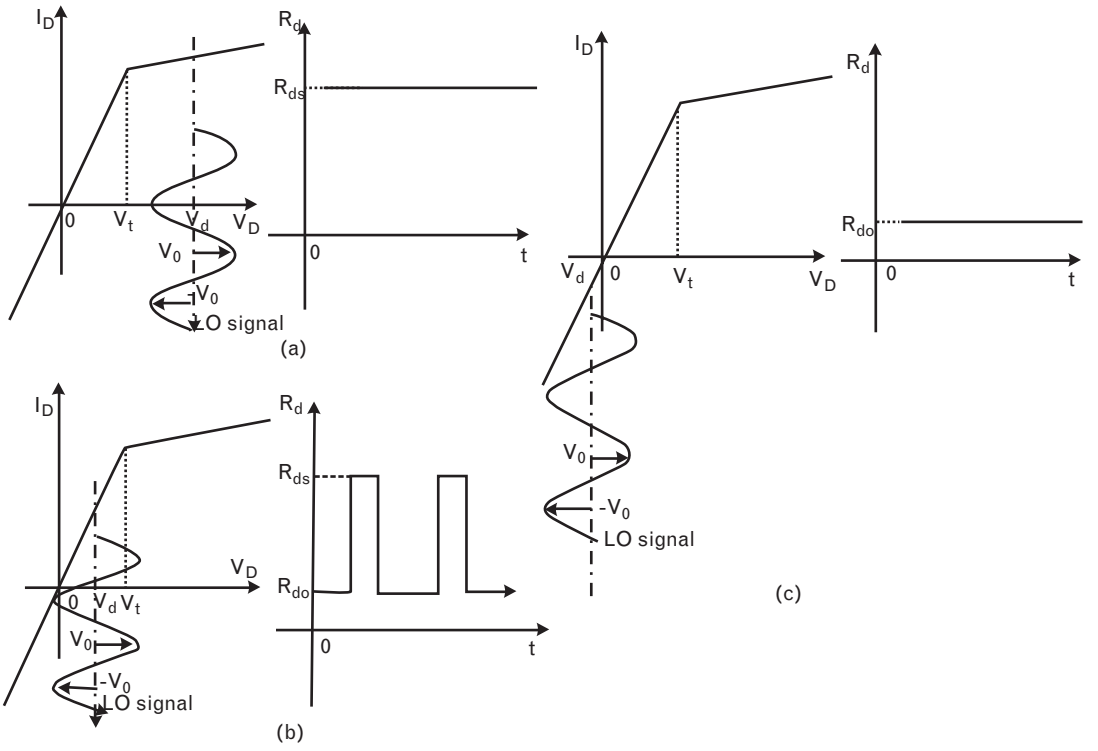


FIGURE 7.47 *Operation of a mixer with drain or collector local oscillator injection: (a) LO bias voltage above the knee, (b) LO bias voltage around the knee, and (c) LO voltage below the knee.*

the maximum swing in the g_1 conductance component at the LO frequency, the local oscillator voltage should swing around the knee of the output I-V curves, between the saturated and linear regions of the transistor. This not only switches the transistor output conductance between high and low values, but also its transconductance since the spacing between different I-V curves is much larger in the saturated region than in the linear region. If the device does not swing through the knee of the curve, the transconductance and output conductance remain either high or low, but do not switch between two significantly different values. FET mixer circuits that attempt to achieve this switching action by driving the drain directly with the LO signal are known as *drain-pumped* mixers.

Figure 7.48 shows this principle deployed in a dual-gate FET, where the drain of the mixer FET1 is driven indirectly by the local-oscillator voltage, via the source of FET2. Thus, the gate of the common-gate amplifier FET2, when driven by the LO, drives its source terminal—the drain of FET1—around the knee of the FET1 I-V curve. As for the other active mixers we have looked at, the applied RF voltage is then applied at a gate

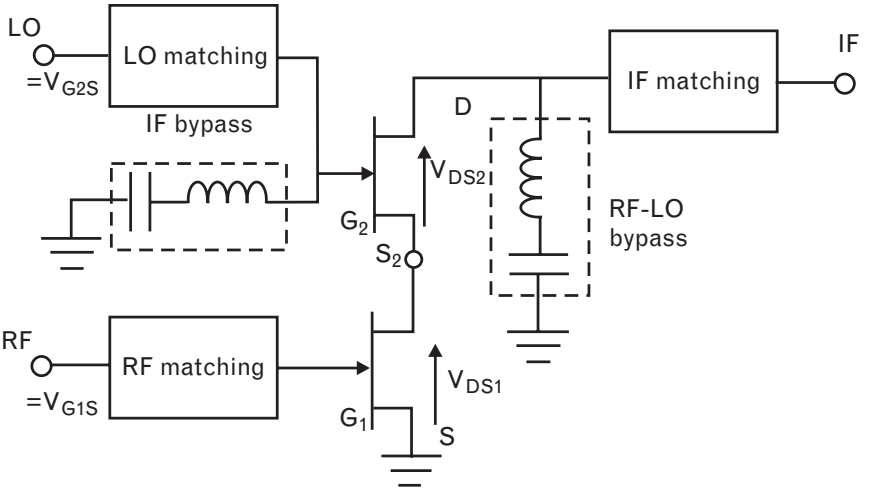


FIGURE 7.48
The general circuit schematic of a dual-gate FET mixer. (From: [2]. © 1998 Artech House, Inc. Reprinted with permission.)

to generate an IF component of current through the variable (trans)conductance, as in (7.3).

Here the RF voltage, applied at the gate of FET1, modulates the drain current primarily through multiplication with the transconductance of FET1, but also with its output conductance. The purist will note that some mixing will, in fact, occur in FET2 as well, because the RF drain current of FET1 will be modulated by the LO in FET2. However, FET2 is in its saturated (amplifier) region where its transconductance is relatively constant and its output behaves as a current source, so this should be a secondary effect.

The LO and RF should in theory be matched at their respective ports. However, the input impedance of the second FET is very high because its source is not grounded, making matching of the LO very difficult. Although the LO VSWR could be improved by crude shunt resistive loading on the gate of FET2, this can significantly reduce the available LO power and will limit the bandwidth since it appears in parallel with the FET2 input capacitance.

The RF and LO currents should be short-circuited at the output drain to minimize distortion through feedback effects, and to keep the IF and RF/LO isolated from each other. The IF should also be short-circuited at the gate of the second FET, since the second FET should operate as a common-gate amplifier to the IF current. As with all common-gate amplifiers, it is important that the gate termination *not* look inductive at any frequency, in order to keep the device stable and avoid oscillation.

The I-V curves of the dual-gate FET appear somewhat confusing because the I-V curves for two FETs are superimposed. Both FETs share the same drain current, thus a common vertical axis. In fact, this is quite an important point, because we could not do this if both FETs were biased in their saturation region, since then both devices would look like current sources connected in series, each attempting to drive the other's (infinite)

output impedance. Consequently, one FET (FET2) will be biased in saturation and the second FET forced to be in its linear region.

The sum of the two drain-source bias voltages across the two FETs is a constant. If, in the example shown, the total bias voltage is 5V, the I-V curve for FET1 can be plotted as in Figure 7.49 in the normal manner, with V_{DS1} increasing from, say, 0V to 5V. The curves for FET2 need to be plotted with an inverse x-axis with V_{DS2} decreasing from 5V to 0V (left to right), so the sum of the two drain-source voltages at any point on the x-axis equals 5V. However, the gate-source voltage for FET2 is an intrinsic voltage V_{G2S2}; the external gate voltage, which equals the LO voltage, is given by

$$V_{LO} = V_{G2S} = V_{G2S2} + V_{DS1} \tag{7.46}$$

When FET2 is redrawn with I-V curves corresponding to the externally applied gate voltage as the control voltage, the curves for the second device appear as in Figure 7.49. The drain-source voltage of FET1 should swing at the LO rate in the area marked "A" in the diagram, along a line of relatively constant V_{G1S} so that maximum drain current modulation is achieved by the LO, as illustrated earlier in Figure 7.47. In fact, V_{G1S} itself is modulated at the RF rate, so this moves the drain current onto marginally different I-V curves of FET1, but this is a small-signal movement compared with the large LO modulation that switches the operating point above and below the knee. To achieve this, the dc bias on the first gate, V_{G1S}, is set close to zero volts, while the gate-source voltage on the second

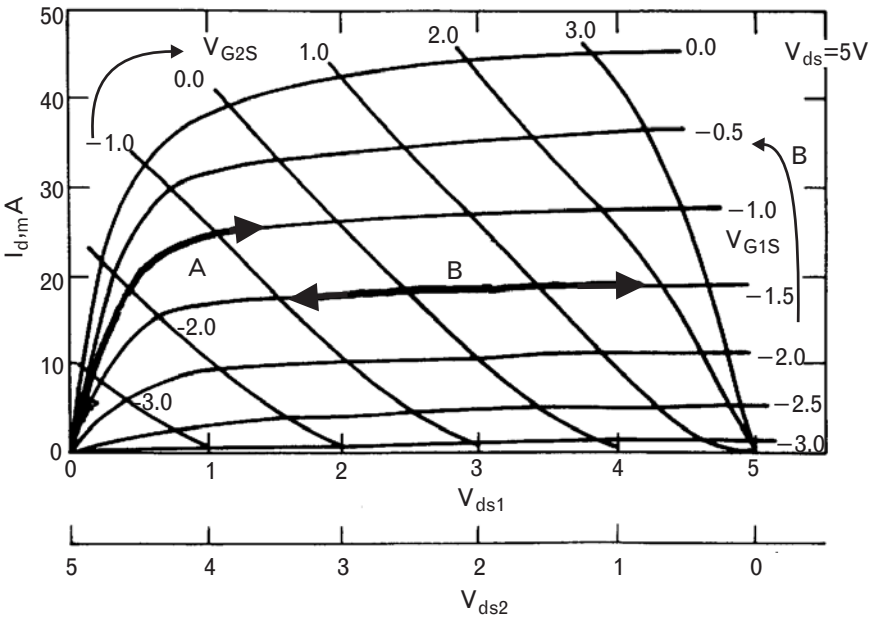


FIGURE 7.49
The I-V curves of the dual gate FET showing the operation of the FET in a region of changing output conductance.

gate swings widely enough to achieve the drain switching. The resulting drain current contains the difference or sum frequency component at the IF and is coupled out the drain terminal of the dual-gate FET.

To the IF, FET2 can be thought of as a common-gate amplifier, buffering the output current from FET1. But unlike a common-gate amplifier, the (variable and often large) output conductance presented by FET1 loads the input and effectively appears in shunt with the output IF load impedance. Thus, IF power is dissipated in FET1 as well as in the output load, and the conversion gain and noise figure of a dual-gate FET mixer are not as good as for a single-ended FET mixer. In a single-ended mixer the output resistance is quite high (since operation is in the flat part of the output I-V curve) and the output current is not shunted by the bottom FET. The LO power requirement is also higher in the dual-gate mixer, because the drain-source nonlinearity used for mixing is less efficient than the gate-source nonlinearity. Nevertheless, the drain of the dual-gate FET output is still a current source, and the mixer can provide some conversion gain provided the load resistance is sufficiently high. As for the single-gate active FET mixer, it is difficult to match this port at the IF because the output impedance of the drain is essentially the FET current source in shunt with a very small output capacitance. Sometimes a high shunt resistance can be used, or else simply the 50-Ω port impedance.

The dual-gate FET mixer requires much lower LO power (of the order of 0 to 5 dBm) than the resistive FET mixer, while its intercept point is only marginally inferior. It also has the interesting characteristic that either the gain, or the intercept point, can be tuned by adjusting the bias voltage on the second gate. As evident from Figure 7.49, this voltage controls the operating point for the output transconductance of the bottom FET, which, since it switches between low and high values, is the principal source of nonlinearity in the device [9].

Because the LO and RF voltages are applied to separate gates, isolation between the two can be 20 dB or greater. The isolation is limited by feedthrough between the gate-drain and gate-source capacitances of the adjacent FETs. As for the resistive FET mixer, the dual-gate FET mixer can be used in both single- and double-balanced configurations to help improve the isolation and remove even spurious responses. Figure 7.50 shows two dual-gate FET mixers driven by both the RF and LO differentially. Because the LO and RF voltages are in phase with each other at each mixer, the IF currents will have the same phase at each drain terminal, so the IF currents will sum at the output and no IF balun is required. This is particularly attractive as the load may be connected directly to the junction of the two drain terminals of each device. Although the IF currents in each device are in phase, the LO and RF currents are out of phase, creating a virtual ground at the junction of the two drains. The requirement to short-circuit the LO and RF at the drain of each dual-gate FET is therefore less stringent than before, since the virtual ground now achieves this. Such

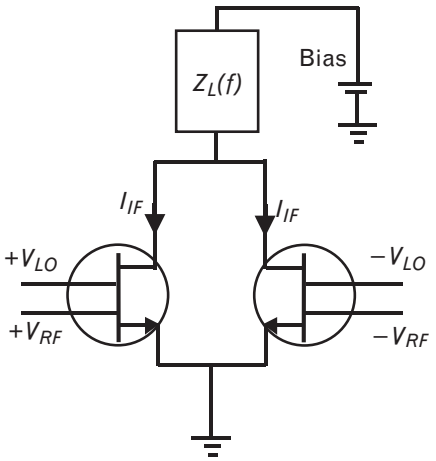


FIGURE 7.50
Dual-gate FET mixer in single-balanced configuration.

a virtual short circuit avoids tuning with narrowband passive matching elements and helps improve both the LO isolation and intermodulation performance.

Figure 7.51 shows a corresponding double-balanced dual-gate FET mixer topology, made of two single-balanced mixers. The RF and LO baluns used to achieve single-ended to differential drive are shown. Because the IF is now also a differential signal, an IF balun is required to properly sum the output currents from the two single-balanced structures. One of these is driven as it would be for single-ended operation as described above. The second is driven with the RF and LO signals out of phase at the two gates of each dual-gate FET. Although, as before, their IF currents sum at the common drain terminal, the IF-pair current is of opposite phase to the IF-pair current in the first single-balanced structure, and an IF balun is required to combine them with the correct phase.

The similarities of this structure to the Gilbert cell mixer are evident, although the mixing processes are quite different. In the dual-gate FET

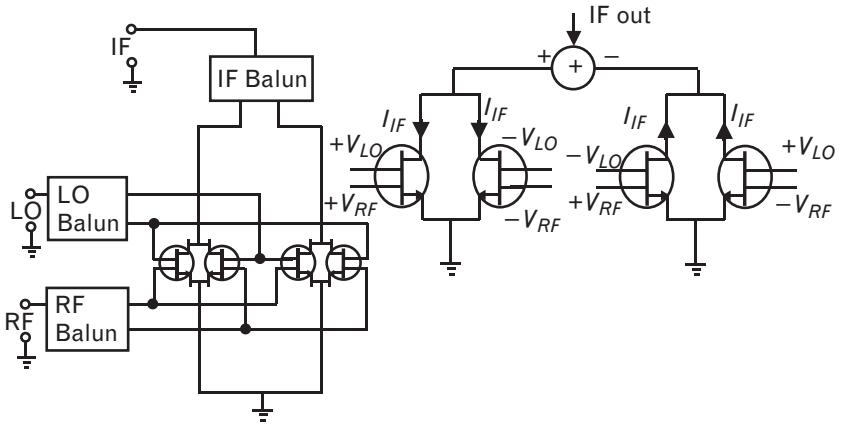


FIGURE 7.51
A double-balanced mixer constructed from four dual-gate FET mixers. A simple equivalent circuit is shown.

mixer, the bottom devices are driven between their linear and saturation regions at the LO rate; while in the Gilbert cell mixer, there should be no LO voltage on the bottom pair, as the top devices are switched between cutoff and saturation. For reasons just noted, the dual-gate mixer requires higher LO power and has lower conversion gain than the Gilbert cell, but will have better intermodulation performance [10].

7.3.4 Comparison of mixers

Table 7.1 compares a single-ended resistive FET mixer with a comparable active, single-ended mixer and a single-ended diode mixer, and presents typical performance values with local-oscillator drive of +10 dBm and an RF frequency of 10 GHz.

All intercept points are with reference to the mixer output. The intermodulation performance of the resistive FET mixer is superior to that of all other mixer types. Even a doubly balanced diode mixer at comparable LO power is unable to offer the same performance as a single-ended resistive FET mixer. Furthermore, because the LO and RF terminals of the single-ended resistive FET mixer are separated, the RF terminal can also be matched for image enhancement. The noise figure of the resistive FET mixer typically equals its conversion loss, because the unbiased device contributes no noise of its own other than its resistive, thermal components. It also avoids the $1/f$ noise problem of its biased counterparts, so is particularly useful for IF frequencies below the noise corner frequency, typically around 1 MHz. However, the noise figure of the resistive FET mixer is worse than its active counterpart, limited by the minimum channel resistance. Active FET mixers have the best noise figure.

For comparison, HEMTs, with their high transconductance and low pinch-off voltage are useful in active mixers when LO power is at a premium. However, their channel resistance is more nonlinear than in a MESFET, and thus, they are a poorer candidates for resistive mixing [2].

Table 7.1 Comparison of Various Mixer Types Showing Various Performance Values

Mixer Type	Conversion Loss/Gain (dB)	IP2 (dBm)	IP3 (dBm)	P_{out}, −1-dB comp.	NF (dB)
Diode	–7.2	9.5	10.5	0	7.7
Resistive FET	–6.5	23.6	21.5	9.1	6.6
Active FET	+6.0	—	16.0	5.0	5.0

Source: [7].

In Table 7.2, we show similar results for an X-band dual-gate FET mixer compared with a resistive FET mixer (different to that in Table 7.1). Intermodulation in the common-source FET of the dual-gate FET causes the poorer intermodulation performance of the dual-gate FET mixer, although its common-gate FET also contributes to the distortion.

When using FET mixers as image-reject mixers, the characteristics of each of the two component mixers should be identical in order to achieve good rejection. This can pose a problem in CMOS, where the FETs are hard to match to each other. At high frequencies, the LO signal needs to be strong in comparison with the threshold voltage of each FET, which may also be quite different. A large LO will help to reduce the flicker ($1/f$) noise contribution of the device, particularly if the mixer is downconverting to a low IF or baseband where its impact is noticeable. These problems become less pronounced at low LO frequencies, where the LO waveform is more square and can drive the FETs with sharp edges during the switching transition [11].

7.4 Frequency multipliers—an overview

There are many similarities between designing frequency multipliers and designing mixers. For this reason, we provide only the briefest of indicators to guide the reader, letting him rely instead on the general principles of amplifiers and mixers we have developed to provide the implicit detail.

Multipliers, mixers, and amplifiers are all driven by external signals, thus amenable to fairly straightforward nonlinear circuit simulation. Both multipliers and mixers have the concept of conversion gain, or loss, between the desired output signal and the fundamental input. The inputs of both multipliers and mixers should be tuned to the fundamental of the input RF signal, and unwanted frequencies should be short-circuited. The output of both circuits should be tuned to the desired IF or harmonic frequency, and unwanted frequencies should be short-circuited. The currents in the devices are driven in a nonlinear region of the I-V curves, frequently class-B, in order to generate a mixing or conversion nonlinearity. Isolation

Table 7.2 Comparison of the Resistive FET Mixer with a Dual Gate Mixer

Mixer	LO Power (dBm)	Conversion Gain (dB)	IP_3 – output (dBm)
Resistive FET	10	< 0	15.3
Dual-gate FET	0	5	13.6

between input and output signals can be achieved through using differential signals and balanced topologies.

Frequency multipliers can be either passive or active in nature. In passive multipliers, a varactor or step recovery diode is frequently used [12], while in active multipliers [2] the design can include any of the transistor classes we have already studied, such as the BJT, FET, and HEMT. The design of frequency multipliers with these devices is not dissimilar to designing a power amplifier, since the source impedance needs to maximize power transfer at the fundamental frequency, and the output load needs to maximize power transfer at the desired harmonic frequency. Ideally, all other unwanted frequencies are terminated in reactive impedances to prevent any loss.

7.4.1 Frequency doublers

In a transistor frequency doubler, the output current produces harmonics through the nonlinear behavior of clipping. This occurs generally whenever the device is biased either class-A in hard conduction or in class-B at device turn-off. For the FET, this occurs for bias voltages of $V_{GS} = 0$ or $V_{GS} = -V_P$, respectively. In the former case, the input voltage waveform will be clipped on positive-going peaks because of forward conduction, and the output current will consist of half-wave rectified current troughs. In the latter case, the device conducts only during the positive-going peaks and the output current consists of half-wave rectified peaks. If the gate is biased midway between zero and pinch-off and the gate is driven sufficiently hard, the gate voltage will clip and clamp symmetrically and the drain current will resemble a square wave. The second-harmonic content will then be relatively small, but the third-harmonic present would then allow frequency tripling.

Figure 7.52 shows the device behavior of the FHX35LG HEMT from Fujitsu when biased in either of these regions. The (measured and simulated) I-V curves for the device are shown in Figure 7.52(a). In region I, corresponding to the class-A bias voltage $V_{GS} = 0$, g_m is high but relatively constant with V_{DS} [Figure 7.52(b)], while the output conductance g_d is highly dependent on both V_{DS} and V_{GS} [Figure 7.52(d, e)]. It follows that the fundamental frequency load impedance should be an open-circuit to generate maximum output voltage swing in order to exploit this dependence. Thomas and Branner [13] show that then the conversion gain for either second or third-harmonic is 15 dB greater in comparison with a short-circuit termination at the fundamental. In region II, corresponding to the class-B bias voltage $V_{GS} = -V_P$, the figure shows that g_m is strongly dependent on V_{GS} [Figure 7.52(c)], while the output conductance g_d is relatively invariant to both V_{DS} and V_{GS} [Figure 7.52(d, e)]. The nature of these nonlinearities confirms that the output current generator is the principal

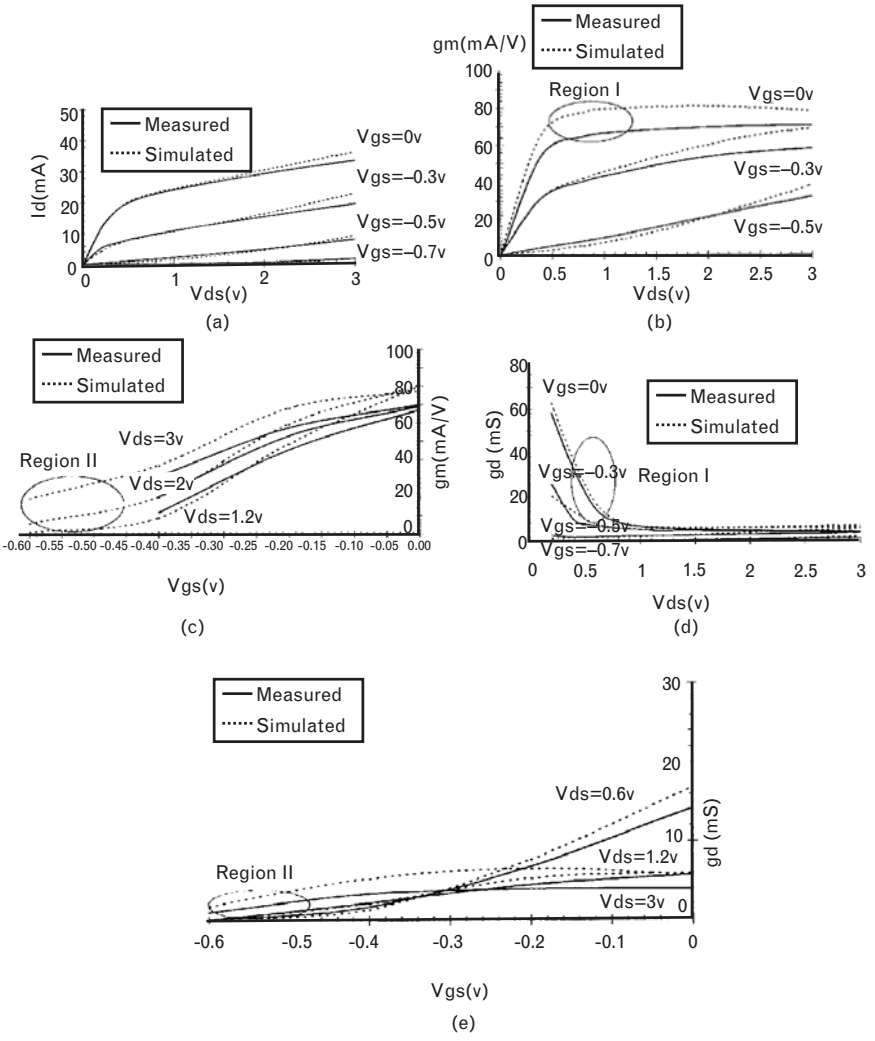


FIGURE 7.52
Measured and simulated curves of the Fujitsu FHX35LG HEMT showing nonlinearities in regions I (class-A) and II (class-B). (a) Output I-V curves: (b) g_m versus V_{ds}, (c) g_m versus V_{gs}, (d) g_d versus V_{ds}, and (e) g_d versus V_{gs}. (From: [13]. © 1996 IEEE. Used with permission.)

source of frequency multiplication in class-B operation. In this case, the fundamental frequency load impedance is less important, but ideally would be a short circuit to generate maximum output current swing. Reference [13] shows a 3-dB reduction in conversion gain as the fundamental load changes from a short circuit to an open circuit.

In both class-A and class-B cases, the input network should be matched to the fundamental frequency and short-circuited at the second-harmonic, for best conversion gain as a frequency doubler. The output network should be matched at the second-harmonic frequency and terminated in the fundamental load described above. However, overall conversion gain on average is several decibels better in region II (class-B) than in region I. The dc efficiency and device reliability are then also considerably improved since the device is biased off rather than in full conduction. As expected from earlier discussions on harmonic behavior, the second-

harmonic output power rises twice as quickly (in dB) as the (suppressed) fundamental output power.

The 3-dB bandwidth of the conversion gain is dependent not only on the transistor itself, but on the input and output matching networks. Since these will be of reasonable Q in order to selectively terminate the desired fundamental and its harmonics, the bandwidth is a compromise with conversion gain. The gain can be recovered using balanced structures [14, 15] that simply parallel the two half-wave rectified outputs of two class-B devices driven 180° out of phase, or that use two 90° couplers to achieve the same result, as shown in Figure 7.53. This creates a full-wave rectified waveform whose second-harmonic component is double that of a single frequency doubler, and therefore achieves a 3-dB power advantage. Using a 180° coupler at the input avoids an output balun since, unlike the balanced mixer, we want the *fundamental* output to subtract and be eliminated. Thus, simply tying the drains of the two FETs together creates a virtual ground at the fundamental, while summing the second harmonic. Using two 90° couplers instead presents good input and output VSWR over a broad range of frequencies.

Such balanced structures can achieve greater than octave bandwidth because the input of each device can be matched over a broad bandwidth, and filtering out the fundamental component at the output is avoided since it is automatically eliminated because of the subtraction that occurs. This gets around the impossible task of trying to simultaneously match and eliminate the fundamental and harmonic frequency components within the same total bandwidth. Avoiding filtering, and separation of the input and output frequencies is a key advantage of balanced structures using transistors.

However, like all balanced structures, the bandwidth is restricted by the phase and amplitude imbalance of the input and output baluns, or couplers. Recently, Piernas et al. [16] have proposed a novel way of accommodating this imbalance by asymmetrically tuning one of the devices (HEMTs in this case), in order to compensate for the phase and amplitude roll-off of the couplers at the band edges. This adjusts the output of the

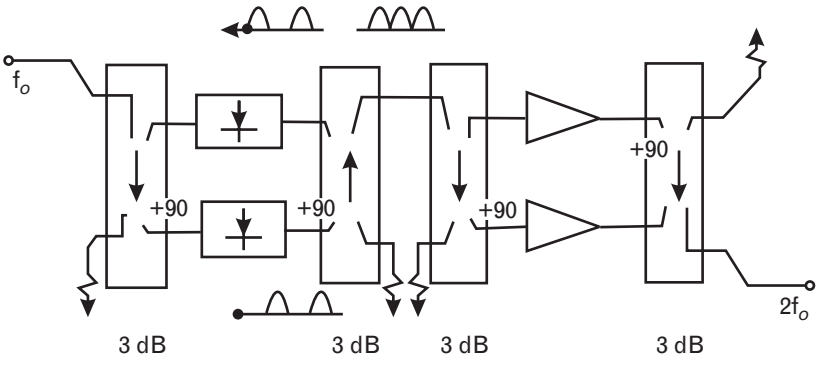


FIGURE 7.53
A balanced frequency doubler constructed from 90° couplers, a rectifying stage, and an amplifying stage. (From: [14]. © 1986 IEEE. Used with permission.)

tuned device relative to the other in order to compensate for the performance of the hybrid couplers. Rejection of the fundamental frequency of up to 40 dB was achieved through tuning the gate-source bias voltage of one of the HEMTs, and expanded the usable bandwidth by 100% at output frequencies as high as 40 GHz.

7.4.2 Arbitrary frequency multiplication

Maas [2] shows how the FET can be biased class-C in order to achieve frequency conversion to an arbitrary *n*th harmonic, by adjusting the conduction angle, or duty cycle, to maximize the desired response. In class-C operation, the duty cycle is less than 50%, and the output current consists of the peaks of sine waves corresponding to a given conduction angle. Figure 7.54 shows the resulting harmonic component in this type of drain current, normalized to the peak current swing, as the duty cycle is increased. The (zero-peak) amplitude of the *n*th harmonic component of the current I_n is given by

$$\begin{aligned} I_n &= I_P \frac{4t_0}{\pi T} \left| \frac{\cos\left(n\pi t_0/T\right)}{1-\left(2nt_0/T\right)^2} \right| \quad n > 0 \\ &= I_P \frac{2t_0}{\pi T} \qquad\qquad\qquad\quad n = 0 \end{aligned} \tag{7.47}$$

where t_0/T is the duty cycle. The figure shows that we should adjust the gate bias voltage and input voltage swing so that the duty cycle is about 30% to maximize the second-harmonic of the output current, or 20% to

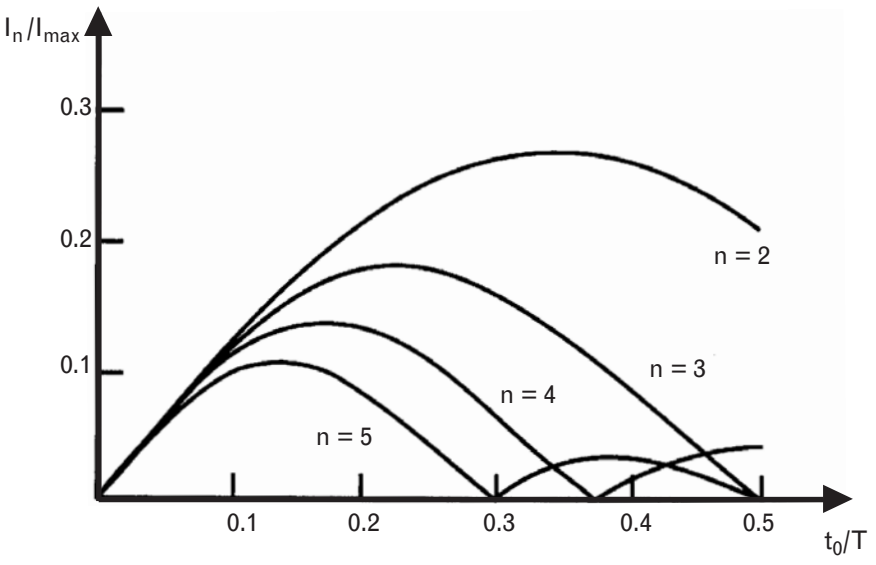


FIGURE 7.54
The magnitude of the nth harmonic current component of a class-C FET (normalized to its peak current swing) as a function of its duty cycle. (From: [2]. © 1998 Artech House, Inc. Reprinted with permission.)

maximize the third-harmonic. The peak of the drain current swing needs to be maximized by driving the input gate voltage as hard as possible, at the same time avoiding forward conduction and being careful that the gate voltage does not swing into breakdown on the reverse half of the cycle. Unfortunately, the conversion gain will be very low with such a short duty cycle as we are forced to use a large voltage swing at the gate.

In principle, the input will be matched to the fundamental frequency and the output to the desired harmonic. The optimum load resistance is typically higher than for an amplifier, as it will simply be the desired (zero-peak) output voltage swing divided by I_n from (7.47) (which is smaller than for an amplifier). However, some thought should also be given to the effect that the output load impedance at the fundamental frequency can have on the feedback through the device [17]. Although, as a general rule we have stated that unwanted frequency components should be short-circuited at their respective ports, the second-order effect of voltage feedback from the drain back to the gate can also have an impact on conversion efficiency. The nonlinear circuit simulations in earlier chapters are an ideal way to experiment with the effect of this and to derive the optimal harmonic terminations for frequency multipliers.

7.5 Problems

1. Because $g(t)$ in (7.1) is time variant (i.e., it is a function of the phase of the LO voltage), even if the amplitude remains constant, both s_{21} and s_{12} are also dependent on the instantaneous phase, hence are also time variant. Here, we define the S-parameters in terms of V_1^+ and V_1^- at the RF frequency, and V_2^+ and V_2^- at the IF frequency. Using explicit time-domain expressions, show that: (a) s_{11} and s_{22} are time invariant; and (b) the expressions derived in Chapter 2 for optimum bilateral match at the RF frequency at the input, Γ_{MS}, and at the IF frequency at the output, Γ_{ML}, are time invariant.

2. Draw the spectra corresponding to the waveforms of Figure 7.4, between dc and 2,500 MHz.

3. The dc forward voltage for the BAT17 mixer diode is 340 mV when the current is 1 mA, and 425 mV when the current is 10 mA. Calculate the sinusoidal LO power levels in dBm required to achieve these two levels of dc current and voltage. What are the corresponding LO impedances at each power level?

4. Using S-parameters for each section of quarter-wave line, show that the input and output impedance of the rat race of Figure 7.12 is 50Ω, assuming each output port is terminated in 50Ω.

5. Why is the conversion gain in (7.26) proportional to the square of the gain-bandwidth product f_T rather than related to it linearly? Answer both intuitively and quantitatively.

6. The active FET mixer in the example in Section 7.3.1.2 achieved an output 1-dB compression point of 11.2 dBm. Reconstruct an FET mixer using an FET of your choice and examine the output IF power as a function of LO power and RF power. Are the shapes of the curves the same as in the example in this chapter? Explain. Try terminating the output LO and RF in an open circuit rather than a short circuit. What happens to the 1-dB compression point, and the conversion gain? Why?

References

[1] Vendelin, G. D., A. M. Pavio, and U.L. Rohde, *Microwave Circuit Design Using Linear and Nonlinear Techniques*, New York: John Wiley & Sons, 1990.

[2] Maas, S., *The RF and Microwave Circuit Design Cookbook*, Norwood, MA: Artech House, 1998.

[3] Mitra, S. G., and S. A. Maas, "A Diode Mixer with Harmonic-Distortion Suppression," *IEEE Microwave and Guided Wave Letters*, Vol. 2, No. 10, October 1992, pp. 417–418.

[4] Liew, Y., and J. Joe, " RF and IF Ports Matching Circuit Synthesis for a Simultaneous Conjugate-Matched Mixer Using Quasi-Linear Analysis," *IEEE Trans. on Microwave Theory and Techniques*, Vol. MTT-50, No. 9, September 2002, pp. 2056–2062.

[5] Gilbert, B., "A Precise Four-Quadrant Multiplier with Subnanosecond Response," *IEEE Journal of Solid State Circuits*, Vol. SC-3, No. 4, December 1968.

[6] Vintola, V., et al., "Variable-Gain Power Amplifier for Mobile WCDMA Applications," *IEEE Trans. on Microwave Theory and Techniques*, Vol. 49, No. 12, December 2001, pp. 2464–2467.

[7] Maas, S., "A GaAs MSEFET Mixer with Very Low Intermodulation," *IEEE Trans. on Microwave Theory and Techniques*, Vol. MTT-35, No. 4, April 1987.

[8] Le, D., and F. Ghannouchi, "Multitone Characterization and Design of FET Resistive Mixers Based on Combined Active Source-Pull/Load-Pull Techniques," *IEEE Trans. on Microwave Theory and Techniques*, Vol. 46, No. 9, September 1998, pp. 1201–1208.

[9] Kim, J., and Y. Kwon, "Intermodulation Analysis of Dual-Gate FET Mixers," *IEEE Trans. on Microwave Theory and Techniques*, Vol. MTT-50, No. 6, June 2002, pp. 1544–1555.

[10] Sullivan, P. J., B. A. Xavier, and W. H. Ku, "Doubly Balanced Dual-Gate CMOS Mixer," *IEEE Journal of Solid State Circuits*, Vol. SC-34, No. 6, June 1999, pp. 878–881.

[11] Abidi, A., "CMOS Wireless Transceivers: The New Wave," *IEEE Communications Magazine*, August 1999, pp. 119–122.

[12] Maas, S. A., *Nonlinear Microwave Circuits*, New York: IEEE Press, 1997.

[13] Thomas, D., Jr., and G. Branner, "Optimization of Active Microwave Frequency Multiplier Performance Utilizing Harmonic Terminating Impedances," *IEEE Trans. on Microwave Theory and Techniques*, Vol. MTT-44, No. 12, December 1996.

[14] Gilmore, R. J., "Design of a Novel FET Frequency Doubler Using a Harmonic Balance Algorithm," *IEEE 1986 International Microwave Symposium Digest*, June 1986.

[15] Gilmore, R. J., "Octave Bandwidth Microwave FET Doubler," *Electronics Letters*, Vol. 21, No. 12, IEE, June 6, 1985.

[16] Piernas, B., et al., "Analysis of Balanced Active Doubler for Broad-Band Operation – The Frequency-Tuning Concept," *IEEE Trans. on Microwave Theory and Techniques*, Vol. MTT-50, No. 4, April 2002, pp. 1120–1126.

[17] Rauscher, C., "High Frequency Doubler Operation of GaAs Field-Effect Transistors," *IEEE Trans. on Microwave Theory and Techniques*, Vol. MTT-31, No. 6, June 1983, pp. 462–473.

CHAPTER 8

Circuits in systems—radio system applications

In this final chapter, we present a brief overview of some modern wireless systems and provide a glimpse into some of the issues concerning how circuits are assembled to provide the required system functionality. We consider some of the trade-offs necessary to obtain higher data rates or reduced bandwidth in a system without loss of signal quality, and the impact these have on their RF design.

In doing this, we first look at the rapidly evolving area of mobile telephony systems, and review the present status of some different wireless telephony standards. Second, we consider some of the issues in how future multipurpose radios might be constructed to be able to cover as many different standards as possible in the one system, and their impact on RF design. Next, we review the design of a cellular system and look at how an integrated circuit design can incorporate many of the considerations we have become familiar with as we have studied component design. Finally, we examine the architecture of a few commercially available chip sets that meet the challenges described in the previous sections.

By necessity, we assume the reader has some understanding of wireless communications fundamentals, since it underpins the systems and standards we are reviewing. Because of the rapid evolution of standards, some of the material we have included on specifications and chip sets will quickly become out of date, so the reader is encouraged to focus mainly on the principles we will illustrate and to use his own research to ensure the content is as current as possible in applying those principles.

8.1 Mobile telephony systems

The analog AMPS system was a first generation cellular telephony system introduced commercially in the early 1980s. Although it started the cellular revolution, it ultimately reached the limits of its capacity because it used analog frequency modulation in an environment where the available

spectrum was band-limited. As the number of users grew, the resulting congestion led to the search for newer systems.

A number of solutions emerged, all centered around digitization of the signal. During the 1990s, mobile systems had explosive growth as costs could be significantly reduced. The real technical breakthrough came about from encoding and compression of voice into efficient spectral densities, more compact than had previously been possible on wired networks. These were exploited by the digital second generation systems for telephony such as GSM, D-AMPS, and CDMA.

More recently, the Internet boom has created a perceived demand for wireless multimedia and much faster data rates. This will be met with the introduction of third generation systems. Recent spectrum licenses for third generation systems have brought tremendous interest and revenues for governments worldwide, although the realization that such networks will cost far more to introduce than originally planned has considerably dampened this enthusiasm at the time of writing.

8.1.1 Second generation mobile systems

The diversity of second generation wireless systems is illustrated in Tables 8.1 and 8.2, which shows some salient features of digital cellular and digital cordless systems, respectively. These systems are described elsewhere in far more detail than necessary here. To the RF engineer, the air interface of these systems is of key interest.

Most digital transmission systems split the available spectrum into a series of frequency channels, each occupied by a carrier frequency; this is known as *frequency division multiple access* (FDMA). Cellular telephony systems usually have a frequency plan so that adjacent frequency

Table 8.1 Relevant RF Features of Second Generation Digital Cellular Technologies

	PDC (Japan)	D-AMPS (IS-54/136) (North America)	N-CDMA (IS-95) (U.S., Korea)	GSM (Rest of world)
Frequency band	800 MHz	800 MHz	800 MHz	900 MHz
Up-banded system	1.5 GHz	1.9 GHz	1.85–1.99 GHz	1.8 and 1.9 GHz
Mobile output power	29–34 dBm	3–38 dBm	23–38 dBm	29–43 dBm
Access method	TDMA	TDMA	Spread spectrum	TDMA
Channel spacing	25 kHz	30 kHz	1.25 MHz	200 kHz
Users per RF carrier	3	3	25–131	8
Modulation	π/4DQPSK	π/4DQPSK	OQPSK	GMSK 0.3

Table 8.2 Key RF Features of Digital Cordless Systems

	PHS	DECT	PACS	CT-2
Frequency	1,895–1,918.1 MHz	1,880–1,900 MHz	UP: 1,850–1,910 DN: 1,930–1,990	864.15–868.05 MHz
Duplex method	TDMA/TDD	TDMA/TDD	TDMA/FDD	FDMA/TDD
Average Tx power (peak)	10 mW (80 mW)	10 mW (250 mW)	100 mW (800 mW)	5 mW (10 mW)
Modulation	π/4QPSK	GFSK	π/4QPSK	BFSK
Channel bit rate	384 Kbps	1,152 Kbps	384 Kbps	72 Kbps
Number of multiplex	4	12	8	1
Carrier spacing	300 kHz	1,728 kHz	300 kHz	100 kHz
TDMA frame	5 ms	10 ms	2.5 ms	2 ms
Radius of service zone	300–500m	100–150m	300m	50–150m

channels cannot exist within the same cell, where a cell is the area served by a base station. Alternate channels, or the channels after the next-closest (or adjacent) channels, can be coallocated in the same cell. This reduces the requirements on receiver selectivity, since the filtering bandwidth to eliminate the unwanted channel can be relaxed. Most systems use some form of FDMA as the initial multiplex method. They then allocate a number of users to the same carrier. This allocation can be made by subdividing the carrier into time slots, known as frames, that are shared between users; this is *time division multiple access* (TDMA). *Code division multiple access* (CDMA) can be used instead. This spreads the carrier over a larger frequency range, trading off the spectral occupancy for output power and maintaining orthogonality between users by multiplying the carrier by different spreading codes for different users.

Table 8.2 shows a few examples of cordless systems, while new variants such as DCT (based on DECT but operating in the 2.4-GHz ISM band) continue to emerge. Because of the localized area of cordless systems, they are able to employ lower output powers (typically milliwatts instead of watts) and have a more generous spectrum spacing than cellular systems. They are simpler in their interaction with the network as the signaling is far less complex because there is no need to support handoff between cells or features like international roaming. In *time-division duplex* (TDD), the receive and transmit signals from each user occupy different time slots on the same carrier frequency, as opposed to *frequency division duplex* (FDD), where they use different frequencies.

8.1.2 Third generation mobile systems

Future and emerging systems are focusing on improving the efficiency with which the spectrum is used, on increasing the bit rate, and reducing the cost of the mobile handset. The prospect of broadband wireless communications to support multiple media (data, voice, video) has yielded a variety of implementation choices that are based on what are known as third generation platforms.

There are five types of third generation platform. Like much of telecommunications verbiage, the nomenclature is confusing, to say the least. TDMA SC single carrier (using EDGE, for Enhanced Data rate for GSM Evolution) and FDMA/TDMA (using DECT) variants are two approved platforms for operators with restricted spectrum to use. Most operators, however, will use one of three variants of *wideband CDMA* (WCDMA) to deploy their 3G networks. The three standards are:

1. WCDMA direct spread/frequency division duplex, or IMT-DS;
2. *Multicarrier* (MC)-CDMA or cdma2000, or IMT-MC;
3. *Time division duplex* (TDD) or time code CDMA, or IMT-TC.

The first two have been allocated spectrum in several bands. Operating band I lies between 1,920 and 1,980 MHz (uplink or mobile transmit) and 2,110 and 2,170 MHz (downlink or mobile receive), with 5-MHz channel bandwidth. Operating band II is allocated the same spectrum as the existing North American PCS variant of CDMA described in Volume I, Chapter 3. These systems use frequency division duplex to separate transmit and receiver channels. TDD-CDMA uses the same channel bandwidth but within a smaller spectral allocation between 1,900 and 1,920 MHz and 2,010 and 2,025 MHz, since transmit and receive channels are not separated by frequency but by a guard period of time. These systems are all being coordinated by different interest groups, but standardized for interoperability through the IMT-2000 initiative of the *International Telecommunication Union* (ITU). Other bands around 1,710 and 1,885 MHz and 2,500 and 2,690 MHz have subsequently been added to the IMT-2000 spectrum.

There are various evolution paths from the existing second generation narrowband GSM and cdmaOne (IS-95) systems to these endpoints, through what have come to be known as 2.5 generation technologies. In general, GSM and digital-AMPS systems will deploy GPRS and/or EDGE upgrades to their existing systems, with WCDMA as the ultimate 3G radio interface. WCDMA will typically have a maximum handset output power of 24 dBm and a chip rate of 3.84 Mcps. Services that can be offered have come to be known as *Universal Mobile Telecommunications Service* (UMTS).

A separate evolution path is mapped out for cdmaOne systems. They will adopt progressively increasing chip rates to support the higher data

rates offered by cdma2000. Although the "cdma2000 1X variant" is available today in the cdmaOne North American PCS spectral band, it uses the same 1.2288-Mcps chip rate for spreading into its 1.23 MHz channel, and offers only limited bit rates.

8.1.2.1 Impact of wideband CDMA on RF design

A CDMA system translates a narrowband sequence of symbols into a signal with artificially wider bandwidth known as a sequence of chips. It multiplies each complex single symbol with a unique, complex spreading code sequence. The length of the multiplying sequence is known as the spreading factor. In wideband CDMA, for instance, speech occupying about 15 kHz at baseband is multiplied by a spreading factor of 128 to occupy the 5-MHz channel. Higher user data rates will require lower spreading factors (ranging from 4 to 512) to achieve the same 3.84-Mcps chip rate.

A CDMA system in effect transmits each coded symbol in a number of redundant ways, and is uniquely identified at the receiver by a correlation process. Detection is achieved in a "rake" receiver by correlating the received signal with a replica of the spreading code. Since all users' codes are orthogonal to each other, there is only one complete sequence with an expected perfect match. Other interfering signals are diminished and appear as noise. A spreading factor of 128 provides about 21 dB ($10\log_{10}128$) of processing gain at the receiver since the amplitude of the despread signal is increased this amount relative to interfering signals, due to the correlation process. If an ultimate signal-to-noise ratio of 5 dB is required after despreading to achieve a reasonable bit-error rate, the RF signal can be buried 16 dB deep in noise and still be detected. Because all CDMA users within a cell occupy the same bandwidth and time slots and are distinguished on the basis of their code, regulation is required to ensure that no single user dominates the "interference." The mobile RF transmit power thus needs to be carefully controlled so that the base station sees the power from all users equally. Furthermore, the much lower minimal detectable signal level at the receiver (spread over a broader bandwidth) also dictates the need to carefully control the isolation between transmitter and receiver.

Multichannel systems that support multiple carriers require the *adjacent channel power* (ACP) to be as small as possible to avoid interference between channels. ACP differs with each modulation format used, since the filtering required for infinitely sharp roll-off of the modulation sidebands is impractical to realize. However, even if such perfect roll-off could be achieved, it would still be negated by spectral regrowth caused by nonlinearities within the receiver or transmitter chain acting on the desired signal. Third- and higher-order intermodulation distortion generates new components that can directly add to the power in the same and adjacent channels.

As we see in Volume I, Chapter 3, and again in Chapter 5 of this volume, systems with high peak-to-average power ratios are particularly

susceptible to spectral regrowth. In order to maintain a reasonable average output power and efficiency, the transmitter amplifier is operated in an area of reasonable average output power but is subject to occasional high instantaneous peak fluctuations of the signal, causing distortion. Such systems require a large linear range to support a given average power. Either the amplifier must be operated backed-off, or linearization techniques as those discussed in Chapter 5 must be used for the transmitter and/or its power amplifiers.

Orthogonal frequency division multiplexing (OFDM) is one modulation technique that has a high peak-to-average power ratio, typically around 10 dB. It transmits multiple modulated subcarriers in parallel in a single channel, each subcarrier occupying a very narrow bandwidth [1]. OFDM has become a preferred modulation method because of its ability to overcome the problem of multipath interference.

Multipath radio propagation causes multiple echoes of the transmitted signal to be received with delay spreads of up to tens of microseconds. For bits rates in the tens of megabits per second, the *intersymbol interference* (ISI) that results can span up to 100 or more data symbols. Historically, the solution to this problem with single-carrier systems has been to use multitap transversal filters at the receiver baseband to adaptively equalize the symbol. At higher data rates, however, the DSP complexity of deinterleaving up to 100 symbols at a rate of tens of megasymbols per second is exorbitant. OFDM allows much simpler equalization and adaptation to interference to be performed in parallel on a number of slowly modulated carrier signals, each occupying a very narrow bandwidth. The inherent redundancy allows decoding even if some of the subcarriers arrive below the noise floor. Because of the compromise OFDM offers between performance in severe multipath environments and signal processing complexity, it has been accepted as the next generation standard for wireless LAN systems and is currently implemented for digital audio and video broadcasting.

Unfortunately, this makes the job of RF design even more demanding, because the sum of a large number of subcarriers, each individually modulated by a scheme such as QPSK or PSK, will have a high peak-to-average power ratio since the carriers will occasionally all add in phase at the same time. Presently, power back-off ratios of 10 dB are not uncommon in such systems in order to comply with the distortion characteristics of the air-interface specification. The associated cost penalty is one of the motivating factors behind new transmitter linearization techniques discussed earlier. This problem also affects modulation schemes related to OFDM, such as multicarrier CDMA (cdma2000). There, each data bit is transmitted in parallel on multiple independent subcarriers, each of which is then modulated by a single chip of the CDMA spreading code. However, that system attempts to reduce the high peak power ratio through an intelligent choice of code, to avoid each of the multiple sinusoidal carriers adding in phase. Then, as more users add together, the peak-to-average ratio actually

decreases as the sum of a large number of signals approaches a Gaussian distribution. Nonetheless, the requirement for linearity while maintaining efficiency and reasonable output powers is still critical to such third generation wireless systems.

8.2 Software-defined radio

As the number of different wireless systems grow, the need for interoperability is being addressed through multimode radios that support multiple standards, for instance both AMPS and GSM or CDMA. Today, such radios use one receiver chain for each standard, and channels are selected using fixed analog-defined channel filters. However, given the absence of a single standard, the ability to reconfigure the radio to each standard on demand is more appealing because of the flexibility and apparent cost advantages it could provide.

The goal of *software-defined radios* (SDRs) is to enable coverage of multiple radio systems with a single handset using common hardware whose configuration is under software control.

In principle, SDR systems could use a single wideband analog stage and convert all channels to and from digital form by a single high-speed *analog-to-digital converter* (ADC) in the receiver, or a *digital-to-analog converter* (DAC) in the transmitter. At the receiver, the desired channel could be selected from the digitized carrier waveform by software-defined channel selection filters within the digital signal processor. Analog filters would still be essential to limit the noise bandwidth, to prevent aliasing, and to limit the bandwidth to prevent spurious signals entering the ADC. Some compromise would be required, since the need for multimode coverage would imply that these filters would need to be kept broadband to cover the entire range of possible input bandwidths. Digital filtering would in principle be used after the ADC to pick out the desired channel component from an array of possible channels that exist within this bandwidth.

Figure 8.1 compares the conventional heterodyne receiver using standard hardware components with a basic SDR configuration.

Although single-band software-defined radios with a narrowband IF channel have been created, it is so far impossible to meet more general requirements covering wider channel bandwidths in multiple bands, at least with today's technology.

8.2.1 RF digital processing

Consider first the possibility of totally omitting even the analog mixer components shown in Figure 8.1(b), so that digitization would occur at RF, as close to the antenna as possible. This would cover all potential

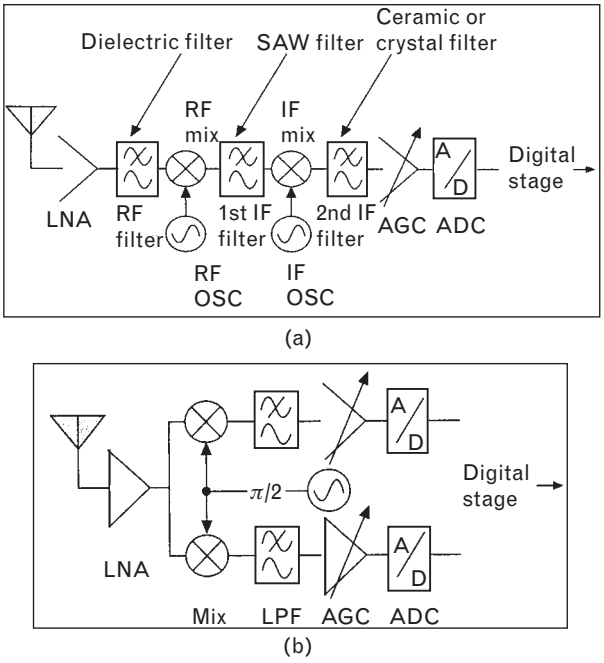


FIGURE 8.1
(a) Traditional heterodyne receiver with standard hardware components; and (b) a basic software-defined radio configuration.

systems, and is in fact the utopian goal of an SDR, since it allows almost complete flexibility. It implies the use of very high speed ADCs and very high processing rates. Such an architecture can be used, for instance, with *global positioning satellite* (GPS) receivers, since the signal is at 1,575.42 MHz and has a narrow bandwidth of 2.046 MHz. However, a GSM receiver, for instance, requires a dynamic range of 97 dB to handle very weak signals in the presence of strong interferers in neighboring channels and must cover an RF bandwidth over tens of MHz. If there is no AGC to assist in handling the variation in signal power, this corresponds to a requirement of 16 bits resolution in the digital representation of the input voltage or current [$20\log(2^{16}) = 96$ dB]. Thus, a GSM signal sampled at RF would require a 16-bit ADC with an analog bandwidth of 900 MHz.

As shown in Figure 8.2, 16-bit ADCs simply do not exist today at operating frequencies much beyond 10 MHz. The AD6645 from Analog Devices is one of the closest in currently available technology, and provides 14-bit samples at 105 megasamples per second. Such CMOS technology still needs to be preceded by a downconverter to IF. However, the sampling limitation is just one requirement, since there is also the requirement that the ADC is sufficiently linear to preserve even the smallest incoming signal in the presence of large interferers. The AD6645 boasts an impressive spurious free dynamic range of 100 dB in this regard.

Once the input has been digitized, DSPs appear attractive for processing the bit stream because apart from channel selection and filtering, they can also provide detection and demodulation of the carrier, fast AGC, companding for speech, frame timing, and error correction, security, and

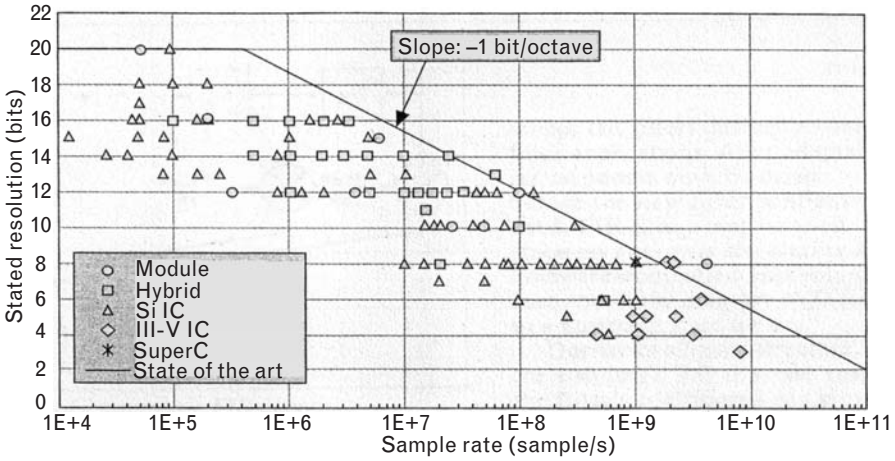


FIGURE 8.2 *A survey of analog to digital converters (speed versus resolution). (From: [2]. © 1999 IEEE. Used with permission.)*

scrambling of the data. ASICs and FPGAs are also well suited for functions such as fast Fourier transform or correlation. However, processing speeds are limited and currently prevent the digital domain from directly encroaching on the traditional radio front end, first because of the trade-off between the sampling rate and resolution, and second because direct computation on the RF signal would require processors with tens of billions of instructions per second compute power. Today's ASICs and DSPs used in processing computationally intensive signals operate two orders of magnitude slower and consume such high power and are so sensitive to timing jitter errors that they are not yet feasible at RF. Although special purpose DSPs for complex and real-time operations such as digital filtering and spreading and despreading of the signal exist, they are still unable to operate directly at RF speeds. Even at baseband, despreading of a multiuser WCDMA signal requires nearly 4 billion complex multiplications per second [3].

Other problems also prevent direct digitization of the RF signal, including the need for a broadband circulator to isolate the receiver from the transmitter (at least in FDD systems), and the potential selectivity requirements of a tunable RF antialiasing filter. These all make digital processing of an entire RF band impractical today and currently restrict the use of software definition to the first IF frequency (and in all probability to a single band) where the digitization is more manageable.

8.2.2 Digital processing of a wideband IF

As just noted, an alternate approach to cover multiple bands and modes is to process the signal digitally at the IF, as illustrated by the system in Figure 8.1(b). Once the analog IF signal is sampled, it can be mixed

digitally with a quadrature LO at the signal's center frequency, to generate baseband I and Q outputs. Known as *digital downconversion* (DDC), digital local oscillators with 100 dB of dynamic range and frequencies approaching 50 MHz are available and enable a narrowband channel to be selected from the multitude that can exist within the wideband IF [2]. Such use of a digital LO followed by a digital lowpass filter can provide excellent selectivity and high dynamic range.

One of the key issues in covering *multiple* system modes is that the IF bandwidths may be quite different for each mode, and the number of narrowband carriers entering what needs to be a wideband IF strip will probably be greater than for a dedicated terminal. For instance, the channel spacing of CMDA is 1,250 kHz while that of GSM is just 200 kHz, so six GSM channels would enter an IF strip wide enough to accommodate the CDMA spectrum. This places harsher requirements on the dynamic range of the components, not only because the noise bandwidth is increased but also because more interfering signals may be present within the IF. The ADC requires a very large dynamic range, since it must be able to detect the desired weak signal in the presence of strong interferers, possibly from unrelated systems. Conversely, the multitude of individual signals on the ADC input can also cause large numbers of low-level distortion terms at the ADC output, increasing the noise floor.

Wideband processing of the IF also stretches available processing power, particularly in handsets. Although within the realms of current technology, a DSP and ADC to achieve billions of complex multiplication steps per second would very quickly consume the battery power. Nevertheless, software-defined signal processing at the IF is seen as the next evolutionary step in SDR, although the extension of multisystem processors and software to the baseband channel modem following traditional analog downconversion will come first.

8.2.3 Digital processing at baseband (direct conversion)

A third approach to SDR is to use a zero IF, which is the direct conversion approach using an analog quadrature downconverter to bring the RF down to dc, with direct I and Q outputs [4]. This architecture is introduced in Volume I, Section 3.1.2. The receiver must now process the full RF spectrum at baseband, so it requires high dynamic range and selectivity. Zero IF appears attractive because digital channel selection is simplified at low frequencies, even if the digital elements still require very high dynamic range and low noise to be able to provide the entire receiver selectivity. Nyquist filters can also be implemented to open the eye diagram and avoid intersymbol interference, and, of course, the image problem is eliminated at dc. The component count and cost are also reduced, since the image filter and IF stages, including the IF VCO and PLL, are not required.

However, because the LO is now at the same frequency as the RF, reradiation of the LO can be a problem. Minimization of internal LO leakage is also important to prevent the generation of dc offsets, since such offsets are not only difficult to cancel, they also reduce the sensitivity since they now fall in-band and can be much stronger than the weaker RF components. The leakage can be controlled through careful process steps and shielding, or through using a harmonic or subharmonic of the LO for mixing. Because the dc offset signal is generated by the mixing of the LO with itself in the second-order term of the transfer function, it can also be controlled by using devices with high *second-order intercept points* (IP2) and maintaining system linearity. Even-order distortion terms are also reduced by using symmetrical topologies with high common-mode rejection (i.e., those with good balance achieved through a virtual ground). In commercial narrowband IF systems using direct conversion architectures, such as in some GSM systems, the dc offsets are also cancelled by using DSP to calibrate the unwanted signal in the GSM idle time slots and then adding compensation. This is not possible with full-duplex systems such as CDMA. However, because CDMA has little signal energy around dc, capacitive coupling is a much simpler approach to overcome the dc problem, although other offsets within the baseband can still be generated from cross-coupling of the transmitter signal or from a strong interferer. The $1/f$ noise also appears at low frequencies and this falls directly on the signal, necessitating the use of devices with low noise-corner frequencies in the receiver.

Direct upconversion in the transmitter, in which the baseband signals are applied to a quadrature modulator whose LO runs at the RF center frequency, is also flexible, reconfigurable, and minimizes the number of components required. However, similar problems exist to those in the receiver. Pulling of the VCO by the high-power modulated transmit signal can result in frequency offset and higher phase noise. As in the receiver, this can also be alleviated by good shielding, on-chip isolation, or mixing on a harmonic or subharmonic of the LO. Also, the dynamic range must now be controlled by principally using the RF amplifiers since there is no longer any IF stage, and this can cause higher power consumption and variable linearity. The transmitter noise floor when the transmitter power is reduced (i.e., set to a low carrier-to-noise ratio) can also cause the phase and amplitude positions of the RF carrier to deviate from their ideal symbol positions in the modulation constellation. This error is quantified by the *error vector magnitude* (EVM), which is a measure of actual signal quality compared with the ideal. It is the ratio of the magnitude of the difference voltage vector between the ideal state and the actual, compared to the voltage magnitude at the ideal state. Any phase or amplitude imbalance in the baseband I/Q signal paths will result in imperfect cancellation of both the LO carrier and the residual (image) sideband, and can be the principal cause of poor EVM.

In spite of these difficulties, the direct conversion architecture is the most common today for SDR, in view of current digital processing limitations of the other approaches.

8.2.4 Transceiver issues associated with software-defined radio

Even the analog RF components of Figure 8.1(b) have special requirements if multiple radio systems are to be managed by common hardware. The LNA will require a very high dynamic range if the input duplexer/preselection filter is removed from the receiver input so that multiple bands can be covered by the same hardware. In that case the LNA itself would need to incorporate linearization and AGC to prevent overload from strong impinging signals (including those from the transmitter), and the mixers would need to incorporate image and spurious rejection.

Kennington [5] quantizes the potential hardware requirements for the full duplex operation of such a radio and gives the example of a typical radio that is required to transmit at 1W (+30 dBm) power levels while maintaining a receiver sensitivity of –110 dBm. Assuming that for common digital modulation formats the signal must be 10 dB higher than the noise floor for adequate detection and low bit error rates, the isolation required of a duplexer would be +30 – (–110) + 10 = 150 dB if the transmit and receive frequencies are identical. Since this is an almost impossible requirement, either TDD must be used with excellent T/R switching arrangements, or a duplex frequency split is necessary, in which the transmit and receive frequencies are different. In the latter case, we would then rely on the selectivity of the lowpass filters in the IF to remove any out-of-band (Rx) signals induced by the transmitter signal in the receive chain.

However, the transmitter can still induce in-band (Rx) spurious components and cause overload. Then the isolation requirement is set by (1) intermodulation distortion in the receiver generated by transmit signal leakage and (2) the leakage from the transmitter noise floor interfering with a weak received signal.

Considering the first of these, suppose for example that the typical receiver third-order intercept point is +30 dBm referred to the input. With 1W transmit output power and a more practically achievable isolation of 50 dB between the transmit signal and the input into the receiver, the leakage power level from the transmitter would be –20 dBm at the receiver input. Any third-order distortion products induced by the leakage would then be at –20 + 2*(–20 – 30) = –120 dBm, falling in-band in the receiver. Since this is now 10 dB below the minimum detectable signal of the receiver, it will not corrupt the bit-error rate. Therefore, this isolation is satisfactory, and achievable. The dynamic range of the ADC required in this case is 100 dB, to differentiate between the signal at –20 dBm and

distortion at –120 dBm, both present in the analog output signal at the input to the ADC. These levels are illustrated in Figure 8.3.

For the second requirement, typical transmitter noise power will be around –75 dBm, so with the receiver noise floor at –120 dBm, the isolation requirement is 45 dB.

We conclude, therefore, that 50 dB isolation between transmitter and receiver is still necessary even if a duplex frequency split is introduced. Kennington suggests a number of solutions for multiband systems that could be used for the duplexer in place of a restrictive analog filter. They include:

1. *Tx/Rx switch:* This prevents transmitting and receiving at exactly the same time instants by switching at distinct Tx and Rx time slots to achieve full duplex operation. The switches can be made very broadband and have the advantage that the solution avoids filtering and places no restriction on the frequency split. In essence, a TDD split is introduced in addition to FDMA.

2. *Circulator:* A three-port circulator can direct received signals from the antenna to the receiver, and direct signals from the transmitter to the antenna. However, typical isolations are limited to tens of dB, and circulators are frequency sensitive.

3. *Cancellation techniques that remove the transmitter signal from the receive path:* An analogy is the feedforward cancellation used for distortion reduction in power amplifiers. Such techniques are an area of research but are complicated by external reflections from the antenna back into the system.

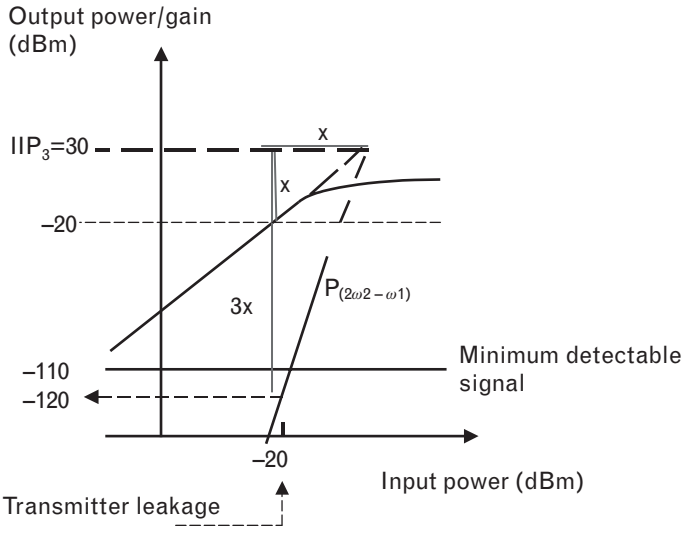


FIGURE 8.3
Calculation of the third-order intermodulation distortion power induced in the receiver by leakage from the transmit signal. The y-axis is the output power referred back to the input after dividing by the system gain.

In summary, the ability to use common hardware to cover all bands and modes in an SDR mobile handset is currently restricted by both analog and digital limits on available technology. As we have seen, there are a number of problems still to be solved, including duplex operation and dynamic range of the LNA; sampling rate, resolution, and linearity of the ADC; processing speed of the DSP; and reducing power consumption.

8.3 A 1.9-GHz radio chip set: design overview

We have covered many issues associated with RF design for wireless systems. In this section, we will review a chip-set design for the 1.9-GHz *personal handyphone system* (PHS), published by McGrath et al. [6], to highlight some of the issues faced by the RF designer in meeting system-level requirements.

8.3.1 The air interface specification for PHS

The PHS system is a Japanese microcellular system with cell sites of approximately 50-m radius. Although labeled a "cordless" system in Table 8.2, the system is in fact more adept than a cordless system since it interfaces with the public network and allows handoff between cells when the user is moving at moderate speeds. The system uses 77 RF channels between 1,895.15 and 1,917.95 MHz (FDMA) and splits each carrier into 5-ms time slots for shared access to each channel (TDMA). The receive and transmit signals also share the same channel frequency, using TDD. The channel spacing is 300 kHz, but within the same cell only alternate channels can be used, spaced 600 kHz apart.

During a transmit burst, the average transmit power at the antenna is +19 dBm (80 mW). The transmitter is switched on and off in bursts. When not transmitting, the transmitter leakage power cannot exceed 80 nW, implying an isolation of 60 dB. The system specification requires that the transmitter should not generate any spurious levels in the RF band that exceed –36 dBm, or spurious levels that exceed –26 dBm out-of-band.

The modulation scheme is $\pi/4$ DQPSK. Binary data is grouped into odd and even bits and differentially coded to generate a succession of quadrature impulses I_K and Q_K, that are passed through a lowpass filter to generate shaped pulses $i(t)$ and $q(t)$. The two signals are combined in a single-sideband modulator that upconverts the paired pulses to create one of four possible phase states on a carrier. Any QPSK system has these modulation states equally spaced around the unit circle. As we see in the section on baseband filters in Volume I, Chapter 8, the transfer characteristic of the baseband filter $H(f)$ determines the occupied bandwidth of the system. For PHS this filter function is defined as

$$H(f) = 1 \text{ for } 0 \le f < \left(\frac{1-\alpha}{2T}\right)$$

$$H(f) = \cos\left[\frac{T}{4\alpha}\left(2\pi|f| - \frac{\pi(1-\alpha)}{T}\right)\right] \text{ for } \left(\frac{1-\alpha}{2T}\right) \le |f| < \left(\frac{1+\alpha}{2T}\right)$$

$$H(f) = 0 \text{ for } \left(\frac{1+\alpha}{2T}\right) \le |f| \tag{8.1}$$

where T is the bit rate and $\alpha = 0.5$. This is a Nyquist filter that avoids intersymbol interference yet limits the occupied bandwidth. It is implemented digitally at baseband in the transmitter, prior to any upconversion.

The filtering causes the transitions between symbols to deviate from the unit circle and to pass closer to zero, so that the modulation envelope is no longer constant, and varies from +2.9 to –11 dB about the average power level at the sampling points. Such a nonconstant envelope modulation format causes spectral regrowth in the transmit power amplifier, since its nonlinearity will change with input power. Higher-order distortion is generated. To limit the allowed regrowth, the system specification requires the distortion power that falls into adjacent channels to be no more than –31 dBm at 600 kHz and –36 dBm at 900 kHz from the carrier. With a transmit power level of +19 dBm, this corresponds to intermodulation distortion levels no higher than –50 to –55 dBc. Given that at the 1-dB compression point the IP3 point is typically 10 dB higher and the third-order distortion products therefore 20 dB lower (–20 dBc), this implies that the power amplifier needs to be very linear, or operated backed-off. However, this would reduce the efficiency of the power amplifier so that the major design trade-off in the transmit chain is between linearity and power-added efficiency.

The air interface specification imposes a sensitivity requirement of –97 dBm on the handset, for a received bit-error rate of 1%. A second specification concerns the impact of strong interfering signals on the sensitivity. It states that in the presence of interferers occupying channels at 600 and 1,200 kHz from the desired channel, the bit error rate should remain at 1% even when the desired signal is as low as –94 dBm and the two interferers are 47 dB higher (i.e., at –47 dBm). Figure 8.4 illustrates this specification.

8.3.2 Component specification

Once the air interface specification is understood, the next step is to deduce the corresponding requirements it imposes on the receiver and transmitter.

8.3.2.1 Transmitter considerations

The goals in the transmitter design are to obtain the required output power, to maximize linearity, to optimize power-added efficiency, and to

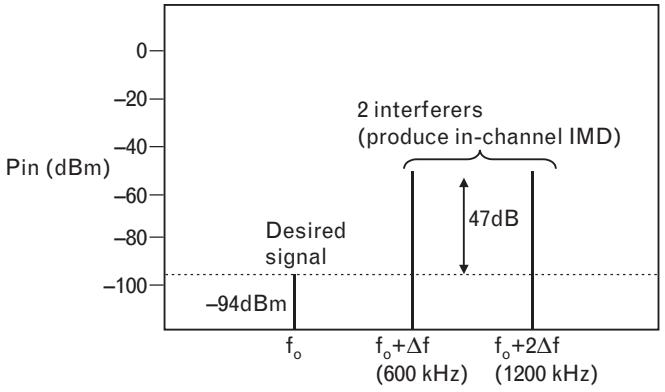


FIGURE 8.4
Air interface specifications for the PHS receiver sensitivity with two interfering tones.

filter unwanted spurious outputs. To obtain maximum efficiency, the power amplifier needs to operate at the highest possible power levels consistent with the ACP specification.

Unwanted spurious outputs are generated by the upconverter in the transmitter. The upconverter is essentially a quadrature mixer, and produces components at frequencies given by $nf_{LO} \pm mf_S$. If direct upconversion from baseband were used, the LO frequency would be in-band around 1,900 MHz. Assuming a LO power around 0 dBm, a minimum of 36-dB isolation would be required from the mixer and RF filtering to achieve the system in-band spurious specification, difficult to achieve in-band.

To alleviate this requirement, the first IF in the transmitter is selected to be either 90 or 240 MHz so the LO will not be in-band. However, spurious frequencies can now arise from other mixing products, and because the transmitter section prior to the power amplifier is likely to be relatively broadband, additional filtering will still be required to remove these spurious components. To filter unwanted out-of-band spurious outputs which cannot exceed –26 dBm, up to 19 – (–26) = 45 dB filtering may be required since the transmit power level is +19 dBm. This isolation will require a physical break in the transmitter package, since otherwise, signal feedthrough within the package itself may ruin any other efforts to eliminate these spurious signals. Isolation between the LO and RF ports will need to be achieved using a balanced mixer topology.

8.3.2.2 Receiver considerations

The goals in the receiver design are to achieve the required sensitivity, to reject spurious products, and to maintain linearity at low supply voltage and current.

For the $\pi/4$ DQPSK modulation format used in PHS, the signal-to-noise ratio prior to detection must be 12 dB or higher to achieve a bit-error rate of 1%. Thus if the minimum detectable signal at the input is required to be –97dBm, the input-referred noise floor should be below –109 dBm.

For a channel spacing of 300 kHz, the final IF bandwidth is set at 225 kHz (53.5 dB). But the input-referred noise floor is given by

$$N_{IN} = kTBF \text{ so}$$
$$F = -109 + 174 - 53.5 = 11.5 \text{ dB} \tag{8.2}$$

If the losses of the filter and microstrip prior to the LNA are estimated at 2.5 dB, the noise figure of the LNA and following components must be better than 9 dB.

When two strong interfering signals at 600 and 1,200 kHz offset are present at the input, their third-order intermodulation product falls in our desired channel. Since the intermodulation distortion from two modulated channels will appear noise-like, it will add to the noise floor. If the distortion is x dB below the noise floor, we can think of the noise floor being raised by an amount $10\log(1 + 10^{-x/10})$ dB. In other words, the distortion appears like noise and adds directly (in milliwatts) to the noise floor. Now to maintain a bit error rate better than 1% at the output, we require

$$\frac{S}{N + D} > 12 \text{ dB} \tag{8.3}$$

where D is the level of the resulting in-channel distortion. The system specification requires a signal as low as –94 dBm to be detected, so with N = –109 dBm, it follows from (8.3) that we require D < –109 dBm as well, assuming the noise and distortion are additive in power (so $D + N$ = –106 dBm and the effective noise floor is increased by 3 dB). From the air interface specification, these conditions apply when the interfering signals are as strong as –47 dBm. Consequently, if the fundamental interferer input power level is at –47 dBm per tone and their third-order distortion is at –109 dBm, the third-order intercept point required is at –47 + 1/2★(–47 – (–109)) = –16 dBm, assuming a 3:1 rise in third-order distortion with input power. Thus, the system specification for the third-order intercept point referred to the input of the receiver is –16 dBm. This calculation is illustrated in Figure 8.5.

8.3.3 Component design

8.3.3.1 IF upconverter (modulator) design

The upconverter accepts Nyquist filtered baseband data signals $i(t)$ and $q(t)$ and is shown in Figure 8.6. The baseband inputs are differential, and drive balanced mixers. The mixers have an in-phase and quadrature LO signal at either 90 or 240 MHz. The structure of the modulator is similar to the image rejection mixer of Chapter 7 and cancels the carrier and one of the sidebands. Perfect cancellation occurs when 0° phase and 0-dB amplitude

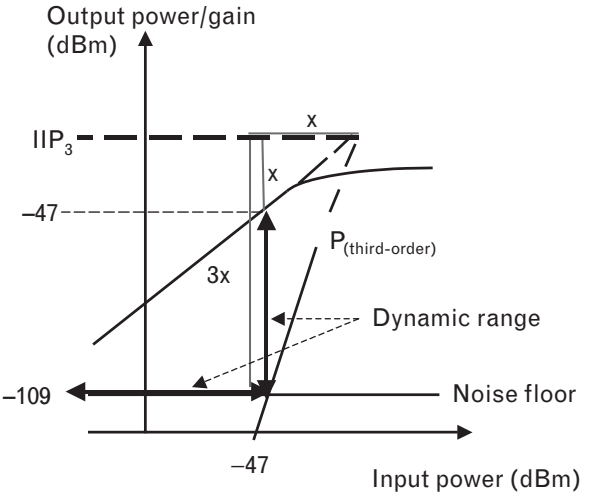


FIGURE 8.5
Derivation of the third-order input intercept point to meet the interferer requirement.

imbalance is obtained. The balun action of the mixer is obtained through the differential action of the combining amplifiers. As a result, the output signal is amplitude-modulated by the total amplitude of the input baseband signal and phase-modulated to an angle equal to the inverse tangent of the quadrature signal over the in-phase signal.

FET quad mixers were used for each balanced mixer, and the SPST switches in Figure 8.6 are series and shunt depletion-mode FETs to achieve the necessary 60-dB isolation between the on and off modes necessary for a time-division duplex transmitter. The IF amplifier uses a two-stage common source network with an off-chip L-C network for matching to 50Ω.

8.3.3.2 Transceiver integrated circuit design

The block diagram of the RF portion of the radio is shown in Figure 8.7. As discussed earlier, the chip set is broken into two in order to insert a bandpass filter to achieve the spurious emission specification required of

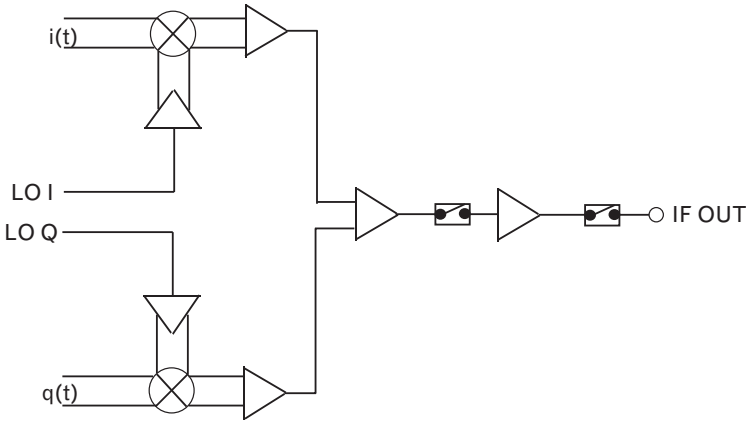


FIGURE 8.6
Block diagram of the QPSK modulator IC. (After: [6].)

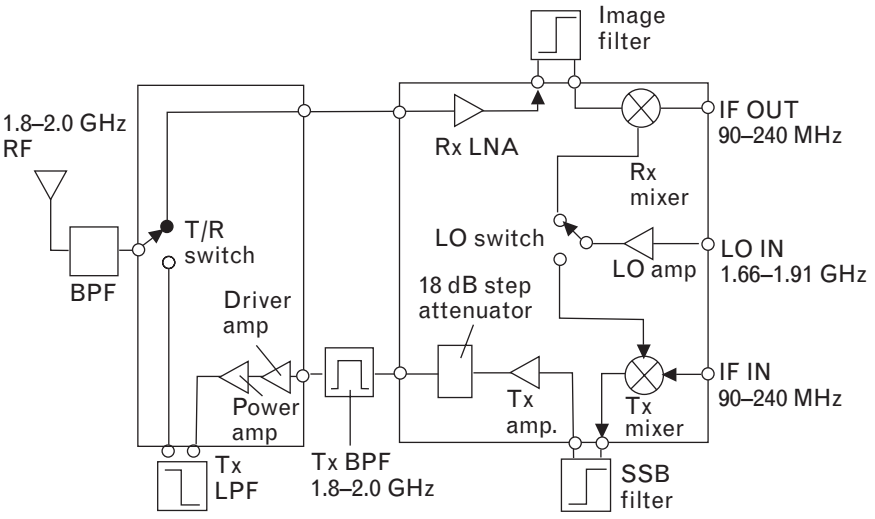


FIGURE 8.7
The RF chip set partition for the 1.9-GHz GaAs PHS radio. (From: [6]. © 1995 IEEE. Used with permission.)

the transmitter. The break shown keeps the high-power components isolated on a separate chip.

Receiver design

The key parameters in the receiver design were to maintain good sensitivity and good dynamic range. The LNA sets the receiver sensitivity and requires good gain to mask the following mixer noise figure. However, the gain must not be so high as to degrade the third-order intercept point. A cascode connection of two FETs was chosen to implement the LNA, as shown in Figure 8.8.

Inductive source-feedback was used in the first FET to create a good 50-Ω input impedance and to move the optimum noise figure close to the same point. The second FET is common-gate, and uses series capacitive feedback in the gate. This and the shunt resistor help to stabilize this common-gate amplifier, which has a tendency to oscillate if there is any gate inductance. Conversely to the dual-gate FET mixer where the bias point is chosen to switch the first cascode device between its linear and

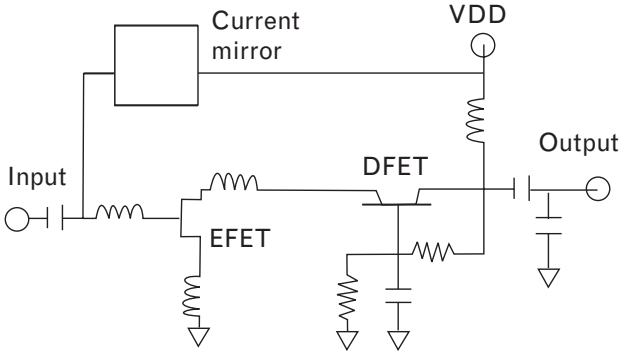


FIGURE 8.8
LNA schematic. (From: [6]. © 1995 IEEE. Used with permission.)

saturation regions for optimal mixing, here an equal split of voltages between the two cascode devices helps to ensure linearity. The LNA is biased at 3V and 2 mA, so its 1-dB compressed point is approximately 3*0.002/2 = 3 mW or 4 dBm, so we can assume its output third-order intercept point is approximately 14 dBm. The LNA achieves a gain of 13 dB and a 2.6-dB noise figure.

The design of the local oscillator was not covered in [6], but we can deduce its phase noise specifications from the system requirements and knowledge of the mixer. Given the mixer uses an active single-ended FET, its LO power requirement will be modest, typically around 0 dBm. The first requirement on its phase noise comes from the need to suppress interfering signals. We have calculated already that an in-band interferer at 600-kHz offset and –47-dBm input power needs to be reduced to a level of at least –109 dBm to preserve the bit-error rate of a desired fundamental signal at –94 dBm (i.e., reduced to –15 dBc). This follows exactly the same reasoning as used earlier to limit the intermodulation distortion produced in the LNA by two interferers. Now, if this interferer itself mixes with the LO phase noise at 600-kHz offset from the LO center frequency, it will be translated directly on top of the desired IF signal. Figure 8.9 illustrates the general principle where power levels are referred to the receiver input. Thus, the integrated phase noise at 600-kHz offset must be less than –47 – 15 = –62 dBc compared with the level used for translation of the main signal by the LO. Since the phase noise at these offsets most likely falls as $1/f^2$ and is higher at one end of the IF bandwidth and lower at the other, we assume (here only, for simplicity to illustrate the point) that this total noise power is just the average derived from a constant spectral density across the entire IF bandwidth noise of 225 kHz (53.5 dB). The phase noise at 600-kHz offset must therefore be better than –62 – 53.5 or –115.5 dBc/Hz.

The second requirement on the LO phase noise can be derived by determining the total phase deviation the noise adds to the signal itself. We

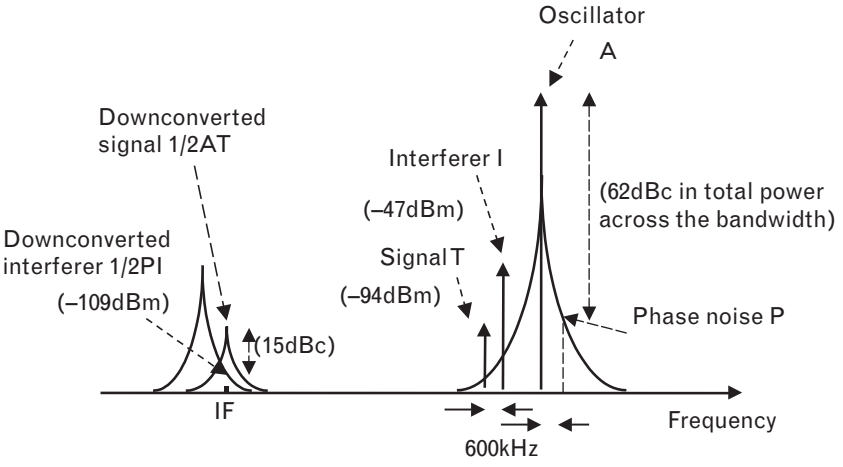


FIGURE 8.9
Principle of LO phase noise calculation showing reciprocal mixing. The LO at A mixes with the desired RF signal at T to produce an IF signal of amplitude 1/2AT. The LO phase noise at P mixes with the interferer at I to produce an IF of amplitude 1/2PI at the same IF frequency.

need to integrate the noise over the channel bandwidth using the equation from Section 6.1.4.4. This is reproduced in (8.4) for convenience below. If we assume a PLL locking bandwidth of 2 kHz, much less than the modulation sidebands, then using –115.5 dBc/Hz at 600 kHz and assuming a phase-noise slope of 20 dB/decade, the noise at 2-kHz offset will be approximately –66 dBc/Hz. If we use the noise power at 20 kHz of –86 dBc/Hz as an average[1] across the IF bandwidth, we obtain

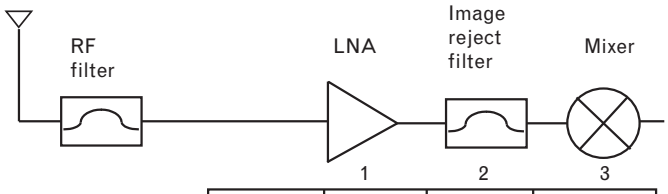
$$\begin{aligned}\left(\Delta\phi_{rms}\right)^2 &= 2\int_{2\,kHz}^{225\,kHz} L\left(f_m\right)df \\ &\approx 2\times 223\times 10^3\times 10^{-8.6} = 1.2\times 10^{-3}\end{aligned} \tag{8.4}$$

so that $\Delta\phi_{rms}$ is 1.9°. This is quite reasonable in a DQPSK system, where the phase separation between adjacent symbols is 45°.

As mentioned, the receiver mixer was designed using a single-ended active FET. The summation of the RF and LO was achieved into the gate by using current sources and a summing circuit. The active FET mixer achieves a conversion gain of 9 dB with an input intercept point of –4 dBm (+5 dBm at the output). The mixer noise figure is 9.5 dB.

Between the LNA and mixer an off-chip image filter is inserted. The image frequency is either 180 or 480 MHz away from the RF signal, depending on selection of the IF frequency, so it can be reasonably filtered without compromising the quality of the RF. If a 2-dB loss is assumed in this component, the RF receiver chain and the component specifications for the LNA, image filter, and mixer are given in Figure 8.10.

The total cascade gain is 20 dB. The intercept point of the amplifier alone referred to the output of the total cascade is 14 dBm – 2 dB + 9 dB =



	1	2	3
G(dB)	13	–2	9
F(dB)	2.6	2	9.5
$IP3_{OUT}$ (dBm)	14	high	5

FIGURE 8.10
Noise figure, gain, and intercept points for the RF receiver chain.

1. The integration should actually be performed on the true phase noise versus the offset frequency plot, but we have assumed some simple averages here to show the principle.

21 dBm; the filter is assumed to have a very high output intercept point; and the mixer output intercept point is 5 dBm. Using

$$\frac{1}{IP_0} = \left[\left(\frac{1}{IP_3\,|\,out}\right) + \left(\frac{1}{IP_2\,|\,out}\right) + \left(\frac{1}{IP_1\,|\,out}\right)\right]$$

to cascade the third-order intercept points of each component referred to the system output, the overall output intercept point of the cascade is approximately 5 dBm, set predominantly by the mixer. Referred to the LNA input, the third-order intercept point is 5 – 20 = –15 dBm, or 5 – 17.5 = –12.5 dBm at the antenna input (allowing for 2.5-dB loss between the antenna and LNA). The system specification calculated above was –16 dBm, so this design allows for around 4-dB production margin.

The cascade noise figure is approximately that of the LNA plus the preceding losses of the RF filter (i.e., 2.6 + 2.5 = 5.1 dB) well within the 11.5-dB specification calculated in (8.2). The margin achieved directly improves the sensitivity of the receiver.

Transmitter design

The mixer in the transmitter is a Gilbert cell upconverter, as described in Chapter 7, which is chosen because of its reasonable linearity and good LO suppression. The allowable level of third-order output distortion from the mixer is estimated at –40 dBc. If the mixer output intercept point is +6 dBm, then this implies the output power needs to be held to –14 dBm, 20 dB lower. The LO power requirement of the Gilbert cell mixer is –8 dBm, quite low, and the LO to RF isolation is just over 20 dB.

In a TDD system, the transmit and receive frequencies are the same. Some care is required to ensure the transmit spurious frequencies are kept low to avoid interfering with other users. These arise predominantly from third-order distortion in the power amplifier, and mixing products in the upconverter. However, upconversion of the LO phase noise can also create in-band noise, and with the transmit spurious power required to be less than –50 dBc in the adjacent channel at 300 kHz, the total LO phase noise at 300 kHz must be lower than –50 dBc/225 kHz, or –103.5 dBc/Hz. Since the transmitter portion provides only broadband filtering across the allocated spectrum, this component is not filtered out. However, since the same LO is used for both transmitter and receiver, the receiver specification is more stringent in this case and thus sets the requirement.

Following the mixer and filter in Figure 8.7 is a driver amplifier with 11-dB gain that needs to deliver 0 dBm of drive to the power amplifier. A step attenuator following the amplifier allows received power levels to be equalized at the base station between different handsets operating at different distances from it.

The main requirements of the power amplifier are to achieve high efficiency and good adjacent channel distortion while operating from a low

operational voltage of 3V. Because of the varying input envelope amplitude, sidelobe regrowth due to the amplifier nonlinearity needs to be controlled.

Figure 8.11 shows the I-V curves of the output FET and the chosen load line. As described in Chapter 5, the load line can be operated at a reduced slope in order to take advantage of the reduced knee voltage and improve the power-added efficiency. Although this reduces the available output power, this is compensated by using a larger device than necessary. A quiescent bias point of 100 mA (20% I_{DSS}) is used. With a knee voltage of 0.5V, the zero-to-peak voltage swing in Figure 8.11 is 2.5V and the zero-to-peak current swing is ideally 100 mA, corresponding to a load line slope of 25Ω. The 1-dB compressed output power is thus approximately $P_O = V_{PEAK}^2/2R_L = 2.5^2/2\star25 = 21$ dBm. The output matching circuit was synthesized using high-Q inductors ($Q = 30$) on chip.

The output power spectrum is shown in Figure 8.12. Measured output power at the 3-V bias was found to be 21 dBm, meeting the system requirement for 19-dBm transmit power at the antenna, after allowing for 2-dB loss in the filter and the transmit-receive switch following the power amplifier.

This chip set was able to be packaged in low-cost surface-mount plastic packages and required only external filters to supplement the RF portion. Yields in excess of 90% have been achieved.

8.4 Integrated system chips: an overview

We have now come full circle. In the first volume, we started with an overview of radio systems and examined some of the requirements of their

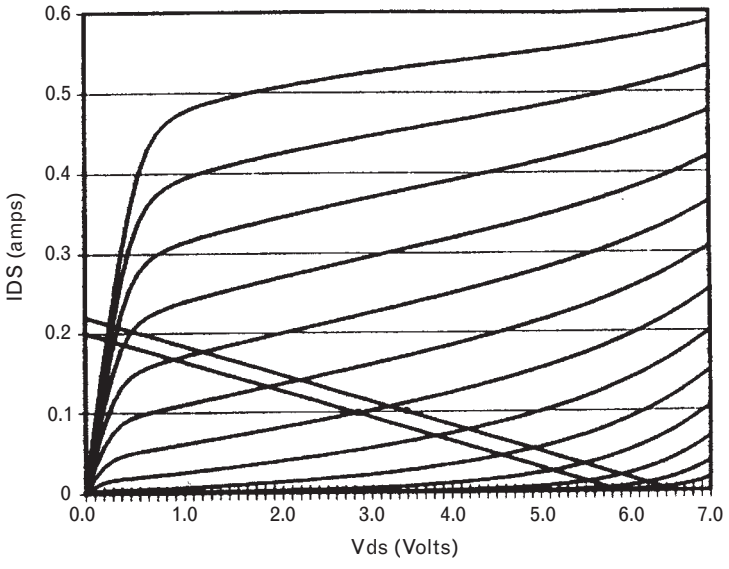


FIGURE 8.11
Load line of the transmitter power amplifier. (From: [6]. © 1995 IEEE. Used with permission.)

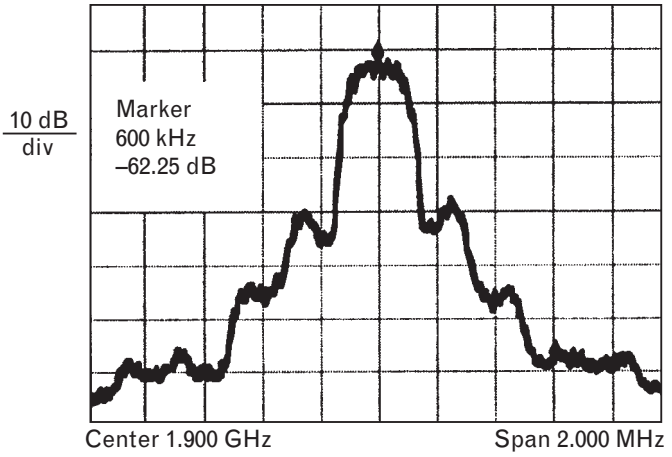


FIGURE 8.12
Output power spectrum of the transmitter power amplifier. (© 1995 IEEE [6]. Used with permission.)

air interface specification—requirements such as occupied spectral bandwidth, transmit power, transmitted spurious products, and the receiver sensitivity. We then related these to the components that make up the system itself and saw that the noise figure, 1-dB and third-order intercept points, linearity, efficiency, and bandwidth of these components all impact in one way or another the final system design. In the ensuing chapters, we looked at the design issues associated with the components themselves and examined the trade-offs in terms of the device technology and circuit topology necessary in order to obtain an optimal solution.

We have now related these components back to the broader context of the wireless system. Much of the systems functionality needed can be achieved using integrated circuits that contain many of the component functions that we have described. Although the IC design fabric can lend itself to circuit topologies different from those we have studied, we now have the tools necessary to migrate to chip design and can examine IC architectures from our vantage point of discrete design.

There are many RF integrated circuits commercially available, and we briefly examine a few of these below. New chip sets are released every year, and the sampling of various integrated circuits below is just that, a sample to illustrate the array. All these chips contain various timing and control functions adapted to the signals required for the particular wireless system in use—for instance, for power management or for signal framing—but we focus here only on the RF aspects.

8.4.1 RF receiver front ends

RF receiver front ends are multifunction chips that perform the receiver functions of RF amplification, mixing, and IF amplification on a single chip. Sometimes, the low-noise RF amplification and RF filtering functions may be performed off-chip, usually because the CMOS technology used to implement the other functions is not the best technology for low-

noise RF amplification, or for high-Q circuits. The local oscillator is sometimes a separate function as well.

The following examples show how multimode/multiband mobile phones are presently implemented. Each band requires separate RF front ends and downconverters to IF, which is possibly a shared IF. An example of such a component is the SA1920 from Philips Semiconductors, an RF front-end intended to cover both the 900- and 1,900-MHz wireless bands. This chip set is designed using a 13-GHz-f_T BiCMOS process and requires a 3.75-V dc supply. Since it is intended for systems as diverse as AMPS, GSM, and PCS, there are no modulation-specific functions on the chip set related to decoding, and because it covers two bands, many of the filtering functions between stages are also off-chip. The low-band section contains a separate LNA and mixer that covers 869- to 960-MHz RF frequencies, with an output IF between 100 and 125 MHz. The LNA has a noise figure of 1.7 dB and 17.5-dB gain; the mixer has 9.5-dB gain and an IIP3 of +5 dBm. When cascaded with a filter between them, the combined noise figure is 2.6 dB.

The high-band section also contains an LNA and an image-reject mixer based on Gilbert cells that operate from 1,805 to 1,990 MHz. The two are internally cascaded, and together achieve 4.2-dB noise figure, 23.5-dB gain, and an IIP3 of –12.5 dBm.

An application circuit is shown in Figure 8.13. The high and low-band LO signals are fed from off-chip (pins 30/31). For the high band, the LO in-phase and quadrature signals are derived by two internal all-pass networks. The IF output signals are internally shifted by 90° and recombined to realize image-rejection. One interesting feature of this chip is that it also contains a separate broadband mixer block for use in the transmitter chain (pin 10). It downconverts the transmitted signal using the same LO as the receiver. This enables the transmitted, downconverted IF channels (pins 2/3) to be used in a closed-loop Cartesian transmitter to improve linearity.

A similar range of products for various RF systems is also available from Maxim Integrated Products. Their MAX2338 is an RF front-end chip intended for dual-mode AMPS/N-CDMA cellular phones, or for other systems such as dual-band GSM. This chip is analyzed at a block-diagram level in Volume I, Chapter 3, and the component itself is shown again in Figure 8.14 within the entire radio. This product uses SiGe technology, which is becoming increasingly popular for RFICs. Like the SA1920, it converts the RF to IF and contains separate LNA sections for both the high (1,930 to 1,990 MHz) and low (869 to 894 MHz) Rx bands, which are fed directly from the duplexer following the antenna. The gain of the LNAs can be switched between several values, allowing adjustment of their IIP3 from +5 to +18 dBm. There is also a "high linearity" mode for higher-power CDMA signals, which increases the power consumption of the chip and is only switched on when necessary. The LNA outputs are fed to off-chip filters for spurious and image rejection, that then drive a broadband

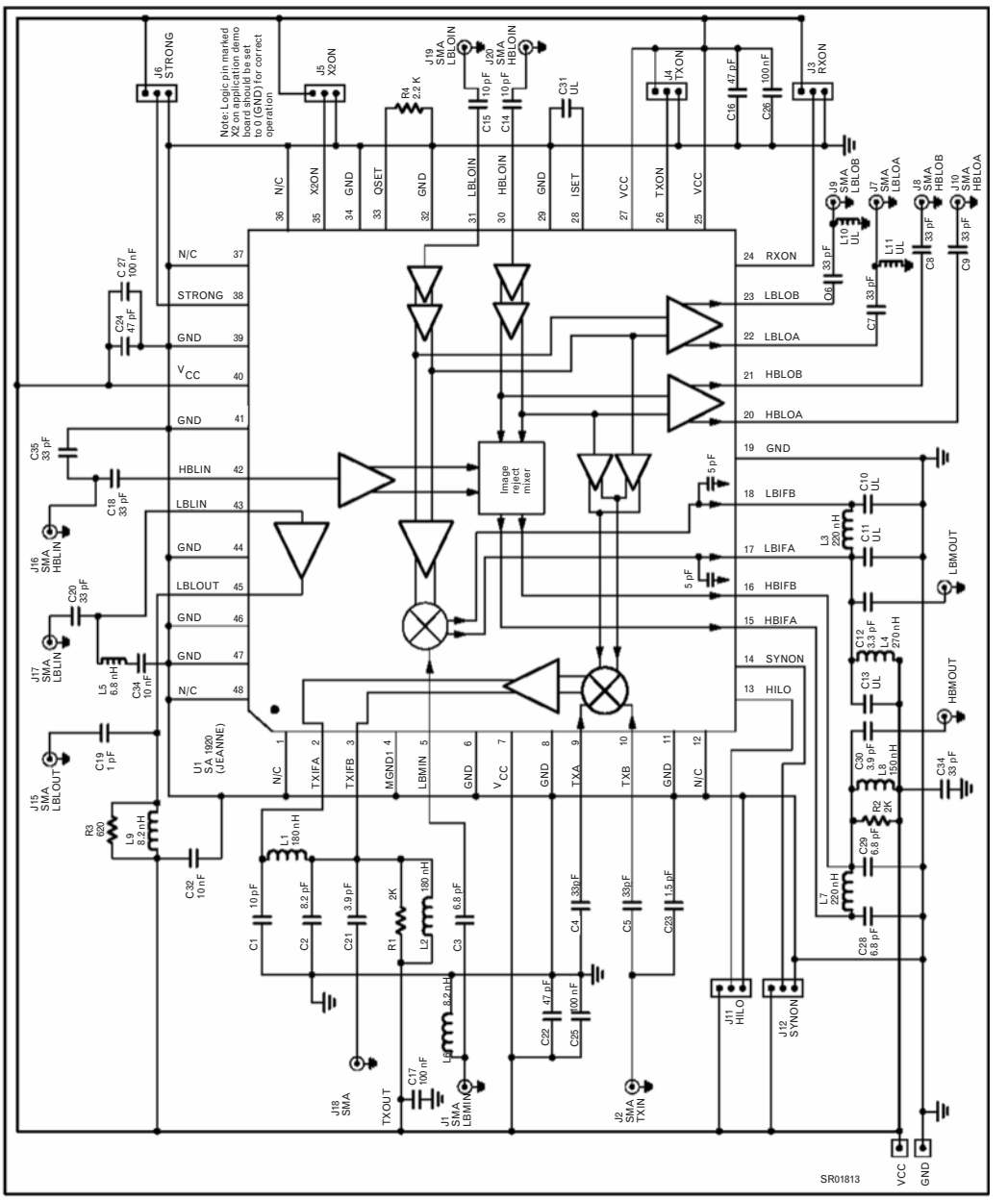

Note: Logic pin marked X20 on application demo board should be set to 0 (GND) for correct operation
Image reject mixer
SR01813
VCC
GND

FIGURE 8.13 *SA1920 dual-band application circuit. (Courtesy Philips Semiconductors.)*

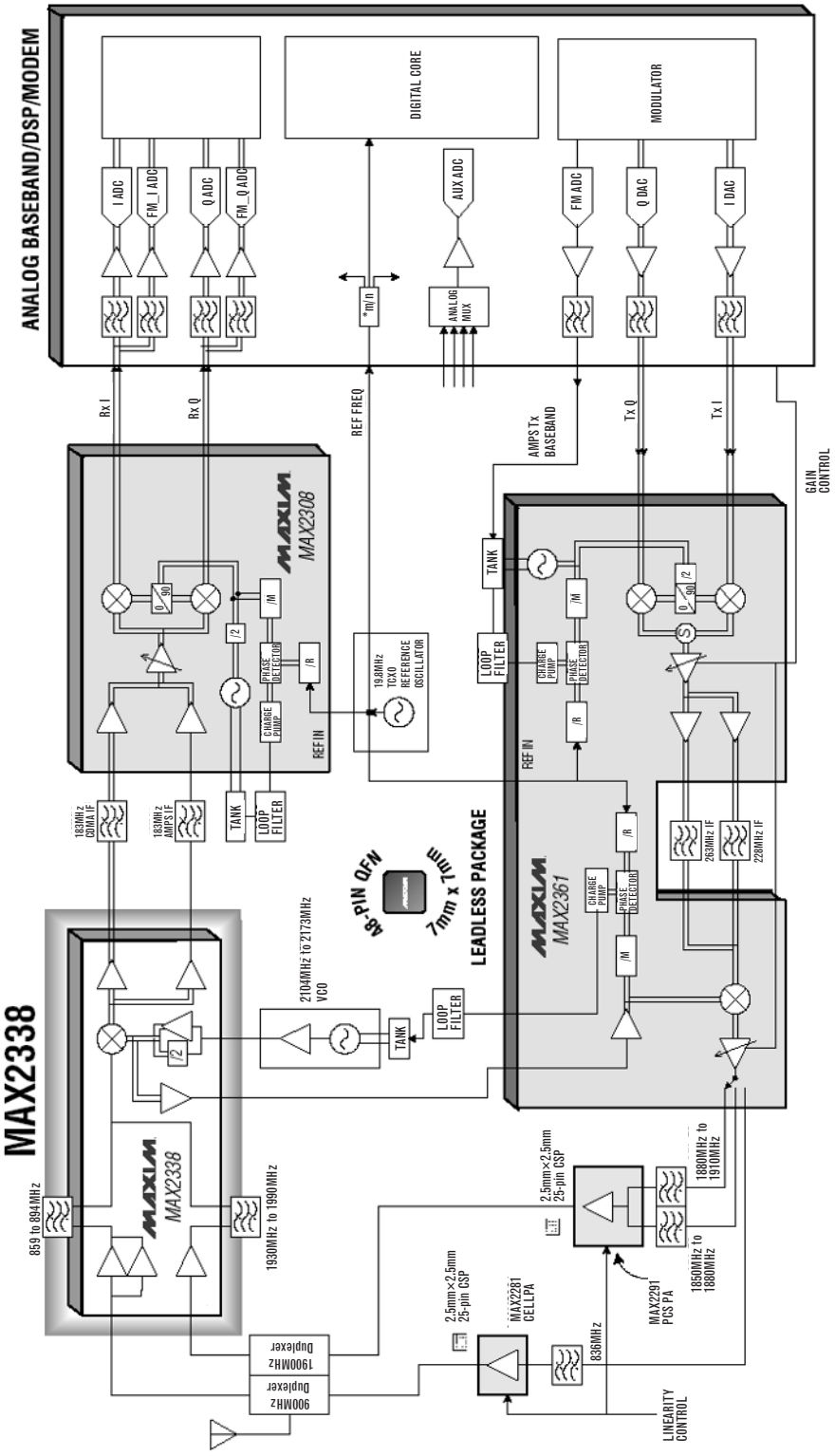


FIGURE 8.14 *PCS phone system block diagram showing the MAX2338 chip and others. (Courtesy Maxim Integrated Products.)*

mixer. The mixer uses an off-chip LO at around 2,150 MHz, which can be divided on-chip by a factor of two for the low band, thus enabling a single IF of around 183 MHz to be used. Separate variable gain IF amplifiers are provided for the two modes since the AMPS and CDMA channel bandwidths are different, and these output the IF signals from the chip for channel filtering. The chip operates with a 3-V dc supply. A planned upgrade, the MAX2538, will add full GPS functionality for position indication, including its VCO.

Another receiver front end is the AD8347 from Analog Devices, whose block diagram is shown in Figure 8.15. It is a broadband (800 to 2,700 MHz) quadrature demodulator, and differs from the previous two in that it has zero IF and is intended for direct conversion receivers that directly drive an analog-digital converter. This chip uses a silicon bipolar process and can also operate from a single 3-V supply rail.

The input RF and LO signals are differential so require an external balun. The RF input (pins 10/11) passes through two stages of variable gain amplifiers and is split into two Gilbert cell mixers. The LO, fed from off-chip (pins 1/28), is split internally into in-phase and quadrature components via polyphase phase-splitters, which are RC networks. Each LO signal then drives one mixer to yield the I and Q components at baseband (pins 8/22), which are also amplified.[2] There is almost 70 dB of gain-control split between the RF and baseband amplifiers to adjust the input intercept point, thus the dynamic range. The gain is controlled by an on-chip baseband power detector and is varied by changing the quiescent collector current in differential NPN transistor pairs, thereby changing their g_m. IF filtering is performed off-chip. The achieved system noise figure is 11 dB at maximum gain, and the IIP3 is a very respectable 11.5 dBm, due to the ability to adjust for minimum gain. A nice feature of this chip set is that it includes a dc offset compensation circuit that nulls out any dc offset component that appears at the output, compared with a reference voltage. As described earlier, such offsets can be problematic in direct-conversion receivers.

8.4.2 RF upconverters and transmitter driver amplifiers

Integrated chip sets are also available for the transmit side of radios, although their degree of integration can be somewhat more restricted because of the higher range of powers required and the tighter specifications on transmit linearity.

The MC13751 from Motorola is an example of a dual-band upmixer and driver amplifier. Its block diagram is shown in Figure 8.16. This is

2. Note that in spite of the apparent similarity in topology, this is not an image-reject mixer because the output signals are not recombined in an output 90° coupler as they would need to be for image rejection. Image rejection is unnecessary here because the IF is at dc.

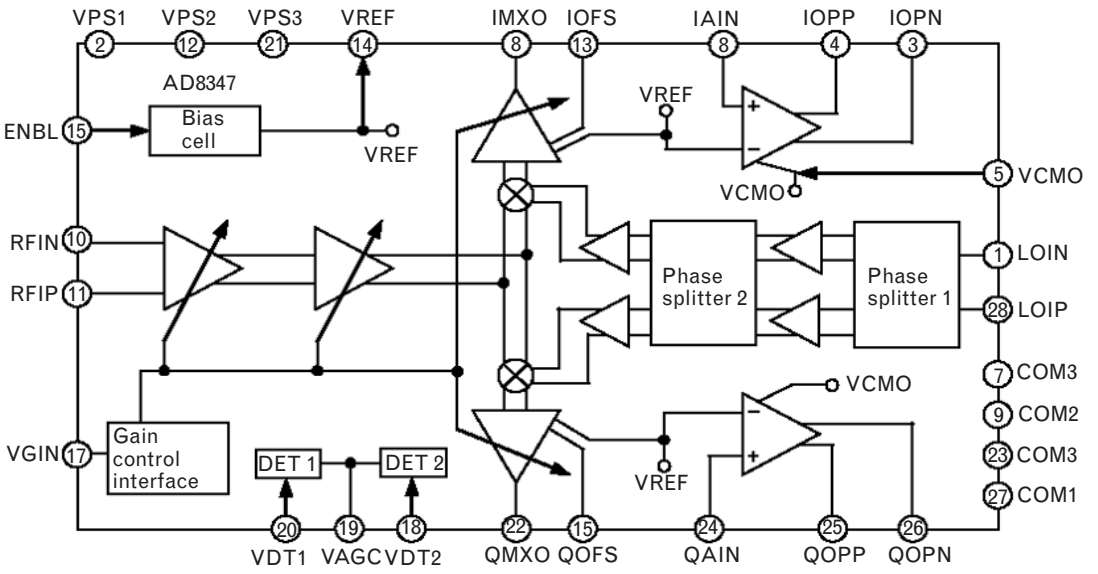
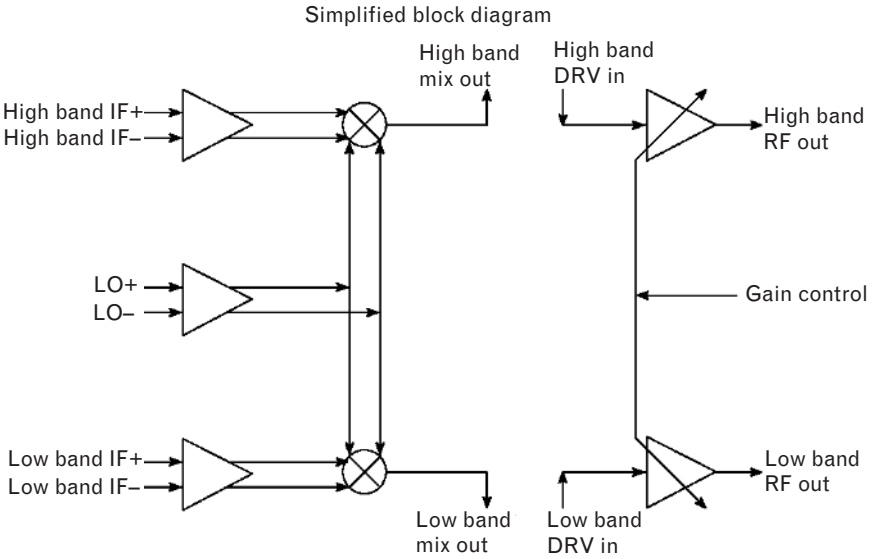


FIGURE 8.15 *Block diagram of the AD8347. (Courtesy Analog Devices.)*

FIGURE 8.16
A simplified block diagram of the MC13751 dual-band upmixer and driver. (Copyright of Motorola, used by permission.)



fabricated in a BiCMOS process, using SiGe. The low band lies between 824 and 849 MHz (corresponding to the AMPS system), and the high band between 1,850 and 1,910 MHz (for the PCS band). The RF impedances are matched to 50Ω. The input IF signals for both bands must lie between 150 and 250 MHz. The LO (off-chip) must be differential at around –10

dBm input power, and lie between 1,002 to 1,029 MHz (low band) and 2,028 to 2,125 MHz (high band). The chip contains separate mixers for upconversion of each band. Their SSB noise figure is a relatively high 11 dB, but this is less of a concern in the transmit chain. These output the upconverted signal for filtering, prior to on-chip amplification to an output power level of around 6 dBm. The level of adjacent channel power and spurious are also part of the chip specification.

The MAX 2361/3/5 series of chips are also transmitter upconverters (baseband to RF), with RF driver amplifiers, each customized for specific cellular systems in the RF frequency bands between 800 to 1,000 MHz and 1,800 to 2,500 MHz, thereby including WCDMA. Architecturally, the series are identical in terms of the functions they provide. The MAX2361 was shown in a PCS phone system in Figure 8.14.

If we take the MAX2363 as an example, it operates from a 3V supply rail and provides +7 dBm maximum output power with –47-dBc APCR (at 5-MHz offset in a 3.84-MHz integration bandwidth), and has 90 dB of power control range in the IF and RF gain stages. The inputs to the chip are the balanced I and Q baseband channels (pins 23-26), which are buffered and upconverted to an IF of 380 MHz by a pair of mixers configured for single-sideband operation (an IQ modulator). The mixers are driven by an on-chip IF LO operating at 760 MHz, although the oscillator resonator is off-chip. The balanced IF output is amplified in a variable gain amplifier and sent off-chip for filtering (pins 16-17). It is then upconverted to RF by a second single-sideband mixer, and amplified in a second variable gain driver amplifier. The IF and RF filtering is kept off-chip in order to allow the use of high-Q, low-loss filters, as is the RF LO. An RF phase-lock loop on chip allows tuning and locking of the LO signal for both the transmitter and receiver signals. A typical application circuit is shown in Figure 8.17.

8.4.3 Transceiver and complete radio solutions

A number of solutions also combine the transmit and receive portions of the system into a single transceiver chip. Such solutions are sometimes preferable to separate chip sets for the receive and transmit sections, although increasing the integration into a single chip reduces the flexibility of partitioning.

The MAX2420 chip from Maxim Integrated Products can be used in a range of applications varying from cordless phones and two-way paging to cellular phones, with RF frequencies from 800 to 1,000 MHz, although it only provides a maximum output power of +2 dBm. Its functional diagram is shown in Figure 8.18.

The receiver path incorporates an adjustable-gain LNA, an image-reject mixer for downconversion, and an IF buffer amplifier. The LNA can be put into a "bypass" mode for best linearity with large signals, or into a class-A mode for small signals. Although the system noise figure is only 4

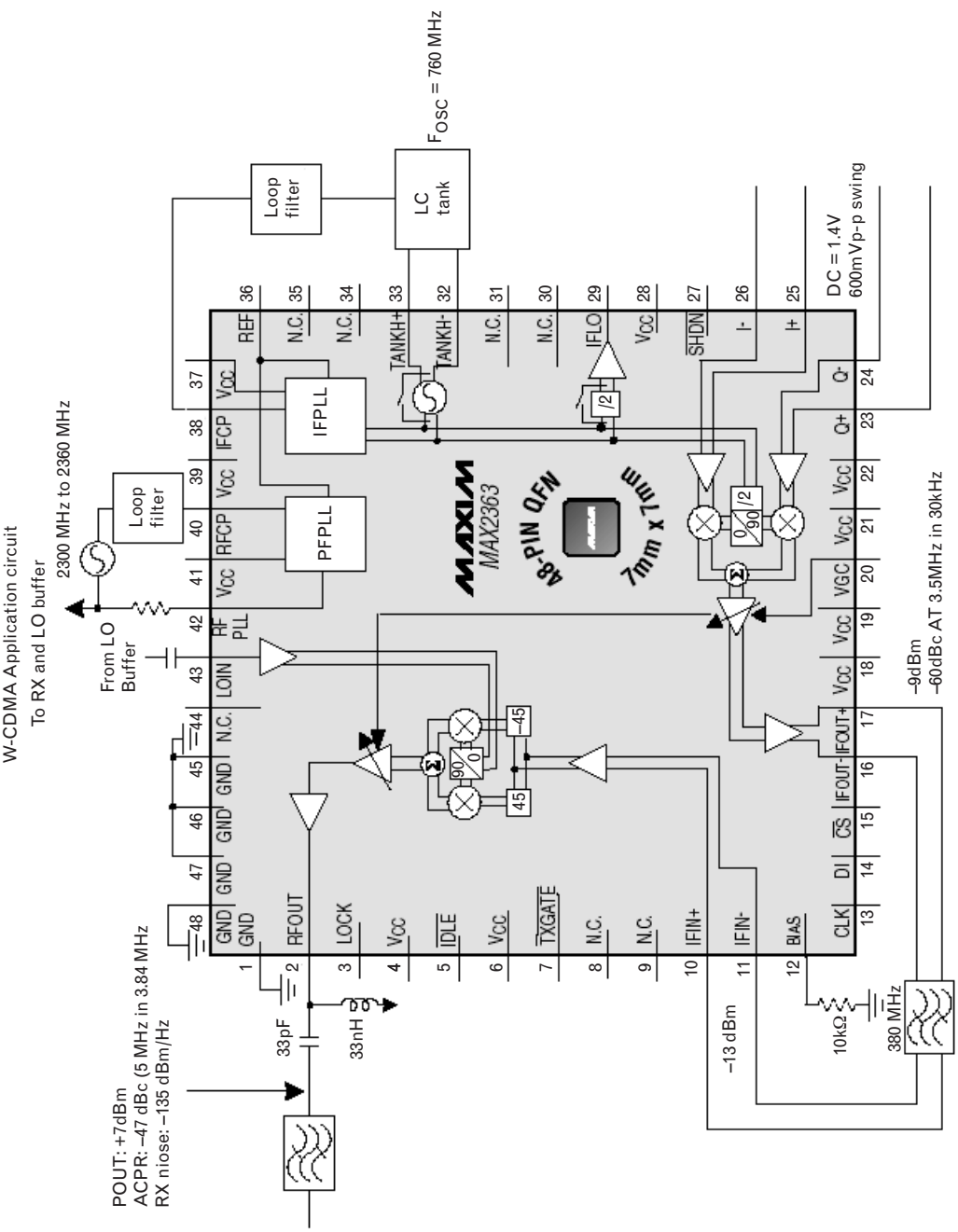


FIGURE 8.17 *A WCDMA application circuit for the MAX2363. (Courtesy Maxim Integrated Products.)*

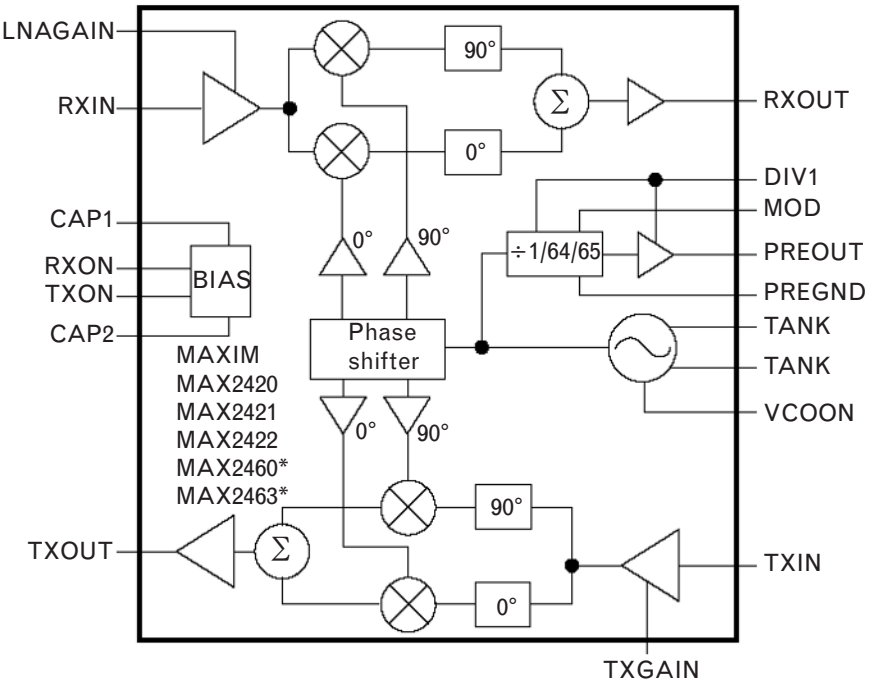


FIGURE 8.18 *Functional diagram of the MAX2420 900-MHz image-reject transceiver. (Courtesy Maxim Integrated Products.)*

dB, the input IP3 varies between –17 dBm to +2 dBm depending on the LNA gain selected, much lower than that from a stand-alone front-end such as the MAX2338. Typical IF frequencies are 10.7, 46, 70, and 110 MHz. Because the image-reject mixers select an upper or lower sideband at their output port, care is needed to ensure that high-side or low-side LO injection is in accordance with that specified.

The transmitter consists of a variable-gain IF amplifier that generates an in-phase and quadrature signal for delivery to the upconverter in an image-reject architecture (IF to RF). The resulting single-sideband is then fed to an RF driver amplifier capable of delivering +2 dBm. Although the power control range is 35 dB, this is still lower than that from the stand-alone transmitter module such as the MAX2361. The LO is derived from an on-chip emitter-coupled pair and uses an external LC tank circuit for varactor tuning.

The MC13190 from Motorola is an example of a low-power ISM band (2.4 GHz) single-chip AM radio, requiring only a DSP or microprocessor for baseband control. It is intended for short-range battery-powered (3V) data links such as remote control, games, and wire replacement applications. A simple block diagram is given in Figure 8.19.

The technology is again a BiCMOS process. It includes an LNA, AM demodulator, and baseband filter in the receiver chain, and a baseband filter, upconverter mixer (AM modulator), and PLL/VCO in the transmitter chain. The typical receiver sensitivity is much less demanding than for

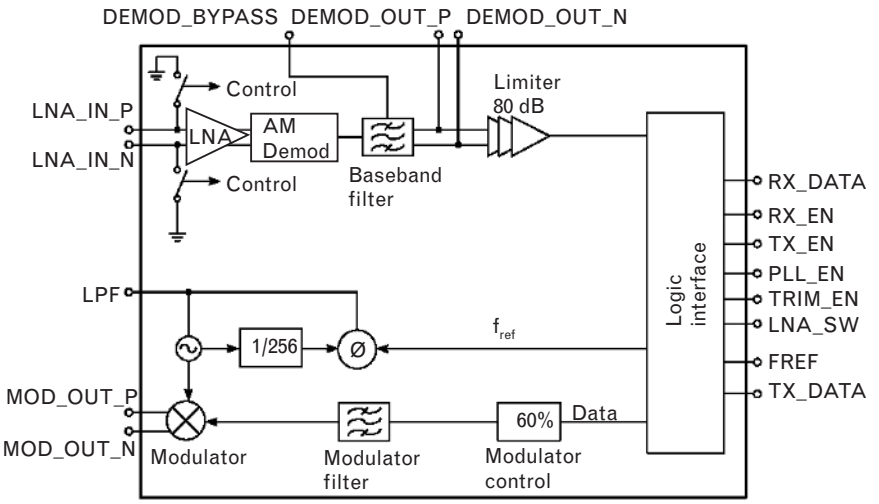


FIGURE 8.19
Simplified block diagram of the MC13190 2.4-GHz low-power transceiver. (Copyright of Motorola, used by permission.)

cellular applications, –68 dBm (when the SNR at the output is 15 dB). The peak transmit power is +8 dBm. A typical bit rate is 5 Mbps at a frequency of 2.442 GHz. The transmit occupied bandwidth is 26 MHz (at the –23 dBc sideband level), and close-in spurious signals are below –36 dBm (at 30-MHz to 1.0-GHz offsets).

The final example of a highly integrated transceiver is the UAA3535HL from Philips Semiconductors, shown with a typical circuit in Figure 8.20. This is a 3-V, low-power module that covers the RF and IF receiver and most of the transmitter requirements of the extended GSM (925 to 960 MHz), DCS (1,805 to 1,880 MHz), and PCS (1,930 to 1,990 MHz) mobile phone systems. Typical current draw is 54 mA when in transmit mode.

In the receiver section, the GSM and DCS/PCS signals must first be filtered and converted to differential form off-chip. Separate LNAs (pins 39/40, 42/43) amplify these signals, and a shared quadrature mixer down-converts them to a near-zero IF of 100 kHz. The mixer provides about 35 dB of image rejection, critical with such a low IF. Channel selection is provided in an integrated bandpass filter, preceded and followed by variable gain amplification, which derive the I and Q channels as differential outputs (pins 7-10). The filter is a fifth-order filter centered around 100 kHz with a bandwidth of 220 kHz. The AGC range is 64 dB.

In the transmitter section, the baseband input I and Q channels (pins 7-10) are first upconverted to a transmit IF of either 45.5 or 91 MHz for the GSM/DCS systems [and 6/7 times that (i.e., 78 MHz) for the PCS system] in a single-sideband mixer (at the bottom of the figure). These IF signals are lowpass filtered and fed to a mixer configured in a "modulation loop" architecture, where it functions as a phase-frequency detector. In this closed-loop architecture, this input transmitter IF signal is compared in

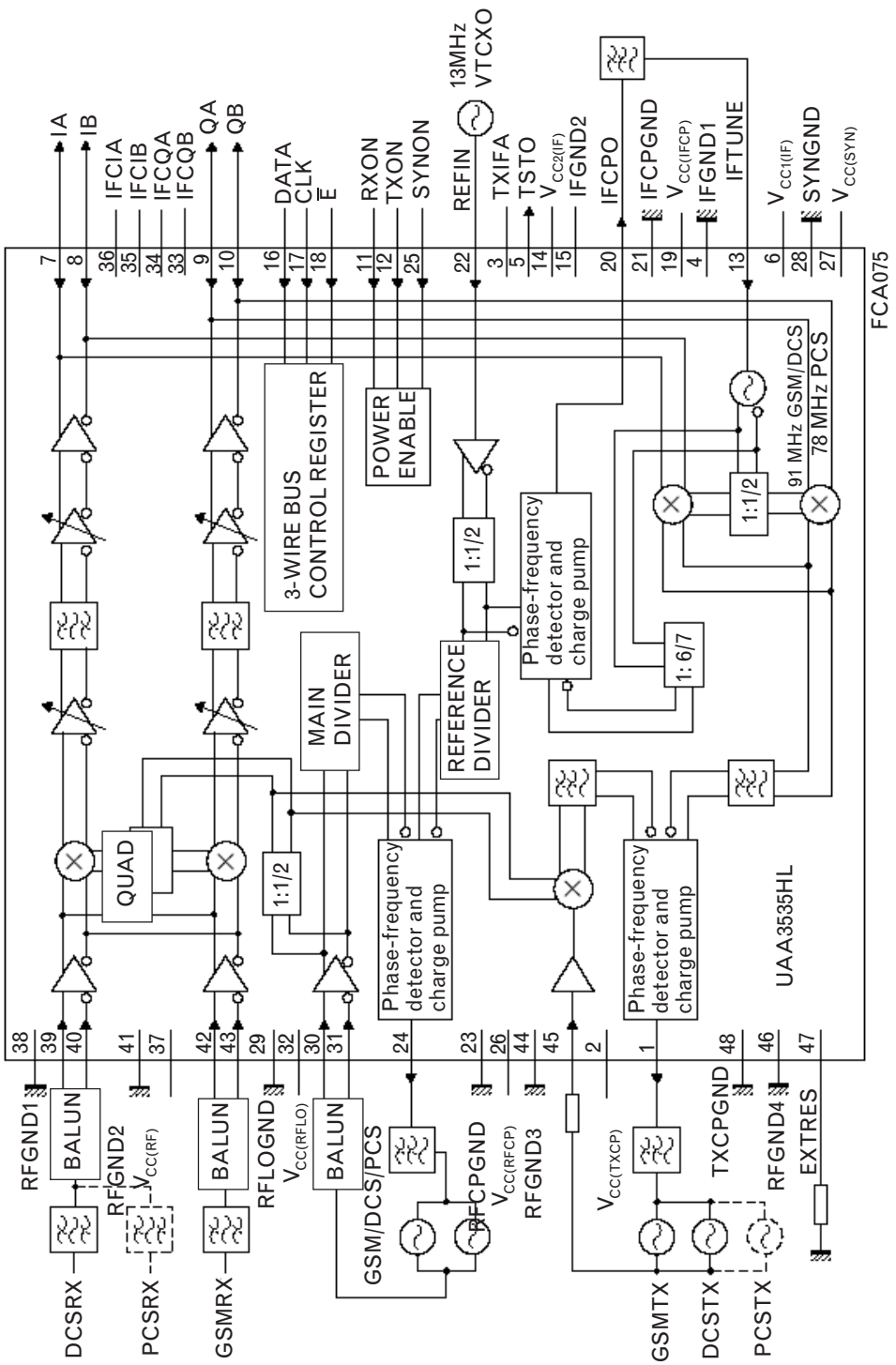


FIGURE 8.20 *Typical application circuit for the UAA3535HL transceiver. (Courtesy Philips Semiconductors.)*

the mixer (labeled "phase-frequency detector and charge pump") with a downconverted version of the RF transmitted signal (at the same IF). This RF signal is generated in off-chip RF VCOs, which are modulated by the phase of the detected transmit IF coming from this mixer. This phase modulation is output at pin 1 in the figure, from the phase-detector mixer. Such a closed-loop system provides excellent phase linearity and very low phase noise, but requires two RF oscillators for each cellular system. One generates the transmitted RF signal itself and would drive external power amplifiers (not shown in the figure); the second provides the RF LO (pins 30/31) to the receiver mixer but also to the downconverting mixer that generates the transmitter IF. Such a system is known as a translational loop architecture, and behaves similarly to a phase-lock loop. It can be used for constant envelope modulation schemes such as GMSK, used in GSM.

An external crystal oscillator (pin 22) provides a reference frequency of 13 MHz for the on-chip IF and RF phase-lock loops. Because of the near-zero receive IF frequency, step programmability of 100 kHz is provided in the PLLs. The PLL loop filters are connected off-chip (at pins 20 and 24 for IF and RF, respectively). An on-chip VCO (controlled at pin 13) is used to produce the transmitter IF, while the external RF VCO referred to above (pins 30/31) covering 1,788 to 2,002 MHz is required as the local oscillator to downconvert the receiver and transmitter signals. A divide-by-two circuit is provided to generate the LO signal needed for the GSM band. The quadrature signals required in the oscillator for the single-sideband transmit modulator and image-reject receiver mixer are generated on-chip.

8.4.4 Power amplifier modules

So far, all of the above modules have used either silicon CMOS or SiGe HBT technology in a bipolar CMOS (BiCMOS) process. This enables the control signals to be integrated with the RF technology and good levels of integration to be achieved at modest cost. Chip sets operating at even 5 GHz (wireless LAN frequencies) are available, and still within the realm of silicon technology.

However, the power amplifier for the transmitter section of such radios is typically implemented separately. This is for several reasons: very demanding linearity requirements that depend on the modulation format of each system; thermal requirements to efficiently dissipate the heat; or efficiency requirements that may dictate a different technology. Only when the power requirements are very low, as in some spread-spectrum systems such as wireless LAN or Bluetooth, is the final transmitter amplifier sometimes integrated with the other transmitter functions. With improvements in material and processing technology, transmit chips with full functionality from baseband to the power amplifier output should eventually be realizable.

An example of a power amplifier chip is the CHP1207-QM from Celeritek. This is a 28.5-dBm power amplifier module for the PCS

frequencies at 1,850 to 1,910 MHz (such as cdmaOne or CDMA2000 1X). It is a cascade of two amplifiers that provide a gain of 27 dB, and is designed using InGaP HBTs. This enables a single supply voltage of 3V to be used. Input and output are matched to 50Ω. High-power, low-power, and shutdown modes are provided via the on-chip bias circuitry. The dc current draw is 590 mA when used in CDMA transmit mode.

A comparable chip for the same application is the MAX2291 from Maxim (also part of Figure 8.14). This chip, operable from a 3V supply rail, uses SiGe HBTs rather than InGaP. It supplies 29.5 dBm output power in the 1,850- to 1,910-MHz frequency range. The transistors operate in class-AB mode, so draw low current in idle mode. The peak current draw is similar to the CHP1207-QM.

The AWT6200 PowerPlexer module from Anadigics is a dual-band (GSM/DCS) power amplifier module. This component integrates in a single package some of the other requirements of a front-end RF system that we have not described in this text, including the T/R switch for the antenna port, transmit lowpass filters for the two bands, and a directional coupler for each band that can be used for control of the output power. It also contains a CMOS controller. The module uses several technologies: InGaP HBTs for the amplifiers, pHEMTs for the T/R switch, and, of course, silicon for the CMOS controller. It is capable of +33 dBm output power for the GSM band and +30 dBm for the PCS bands, at efficiencies of 40% and 30%, respectively. Its functional block diagram is given in Figure 8.21.

8.5 Conclusion

The systems above could only have been imagined a decade ago, when CMOS was not viable above 1 GHz. The trends are clearly towards mixed-mode circuits with high levels of integration combining digital and analog functions, towards digitization of the receiver functions, and towards more complex, closed-loop architectures that exploit these higher levels of integration. We often joke that these advances will put the RF engineer out of work; but then again, maybe digital engineers have the same fears as they see their sphere of influence move higher in frequency and they need to share some of the same tools as their RF cousin!

From here, your next step might be to build on the foundations we have covered in these chapters, and move to IC design, or into system design. The tools we have covered in this book are the fundamental building blocks you need to proceed. We have covered discrete design in detail, and applied the concepts of impedance matching, amplification, oscillation, and frequency conversion in the context of achieving an overall system specification. Although technology will certainly be different a decade

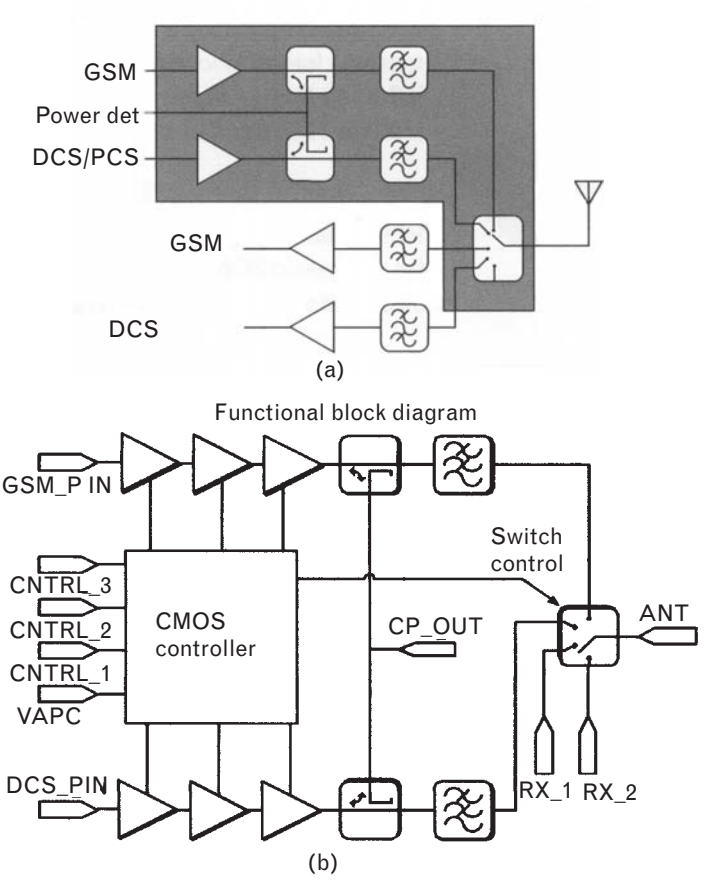


FIGURE 8.21
Functional block diagram of the PowerPlexer transmitter module: (a) function and (b) pin-out. (Courtesy ANADIGICS.)

from now, and the tools more advanced, these same fundamentals will continue to apply whatever the fabric.

REFERENCES

[1] Falconer, D., et al., "Frequency Domain Equalization for Single-Carrier Broadband Wireless Systems," *IEEE Communications Magazine*, Vol. 40, No. 4, April 2002, pp. 58–66.

[2] *IEEE Communications Magazine*, February 1999, special issue on software-defined radio.

[3] Lange, K., G. Blanke, and R. Rifaat, "A Software Solution for Chip Rate Processing in CDMA Wireless Infrastructure," *IEEE Communications Magazine*, Vol. 40, No. 2, February 2002, pp. 163–167.

[4] Loke, A., and F. Ali, "Direct Conversion Radio for Digital Mobile Phones – Design Issues, Status, and Trends," *IEEE Trans. on Microwave Theory and Techniques*, Vol. 50, No. 11, November 2002, pp. 2422–2435.

[5] Kennington, P. B., *Electronics and Communication Engineering Journal*, April 1999.

[6] McGrath, F., et al., "A 1.9-GHz Chip Set for the Personal Handyphone System," *IEEE Trans. on Microwave Theory and Techniques*, Vol. 43, No. 7, July 1995, pp. 1733–1744.

Appendix

Summary of Basic Formulas – 1

(Z_0 is the characteristic impedance and Y_0 is the characteristic admittance)

Reactance

$$X_L = 6.28 f_{GHz} L_{nH}$$

$$X_C = \frac{159}{f_{GHz} C_{pF}}$$

Conductance $$G = \frac{1}{R} \text{ and } g = \frac{1}{r}$$

Susceptance $$B = \frac{1}{X} \text{ and } b = \frac{1}{x}$$

Impedance $$Z = R \pm jX = \frac{1}{Y} = Z_0 \left(\frac{1+\Gamma}{1-\Gamma} \right) \text{and } z = \frac{Z}{Z_0}$$

Admittance $$Y = G \pm jB = \frac{1}{Z} = Y_0 \left(\frac{1-\Gamma}{1+\Gamma} \right) \text{ and } y = \frac{Y}{Y_0}$$

Reflection coefficient $$\Gamma = \frac{Z - Z_0}{Z + Z_0} = \frac{Y_0 - Y}{Y_0 + Y} = \frac{VSWR - 1}{VSWR + 1} = \frac{z - 1}{z + 1}$$

Voltage standing wave ratio $$VSWR = \frac{1+|\Gamma|}{1-|\Gamma|} = \frac{R_{LARGER}}{R_{SMALLER}}$$

Return loss $$RL = -20 \log |\Gamma| = -20 \log \left| \frac{Z - Z_0}{Z + Z_0} \right|$$

Mismatch loss $\left(\Gamma_S = 0, \Gamma_L \neq 0\right)$...

$$ML = -10\log\left(1 - \left|\Gamma_L\right|^2\right) = -10\log\left(1 - \left|\frac{Z_L - Z_0}{Z_L + Z_0}\right|^2\right)$$

$$= -10\log\left[1 - \left(\frac{VSWR - 1}{VSWR + 1}\right)^2\right]$$

Mismatch loss $\left(\Gamma_S \neq 0, \Gamma_L \neq 0\right)$

$$ML = -10\log\left[\frac{\left(1 - \left|\Gamma_S\right|^2\right)\left(1 - \left|\Gamma_L\right|^2\right)}{\left|1 - \Gamma_L\Gamma_S\right|}\right]$$

Wavelength in free air $\lambda = \frac{c}{f} \approx \frac{3\left(10^8\ \text{m}\right)}{f_{Hz}} \approx \frac{30\ \text{cm}}{f_{GHz}} \approx \frac{11.8\ \text{in}}{f_{GHz}}$

Conversion to decibels $dB = 20\log\frac{v_2}{v_1} = 20\log\frac{i_2}{i_1} = 10\log\frac{P_2}{P_1}$

Noise factor $F = \frac{P_{no}}{G_A P_{ni}} = \frac{SNR_{IN}}{SNR_{OUT}} = \frac{T_e}{T_o} - 1 \quad \left[T_0 = 293\text{K}\right]$

Noise figure $NF = 10\log F$

Cascade noise factor $F = F_1 + \frac{F_2 - 1}{G_{A1}} + \frac{F_3 - 1}{G_{A1}G_{A2}} + \ldots$

Summary of Basic Formulas – 2

Unnormalized Form		Normalized Form (using $Z_0 = 50\Omega$ and $Y_0 = 0.02S$)
Ideal lumped component reactance and susceptance of inductors:		
$X_L = 6.283 f_{GHz} L_{nH}$	$[+j]$	$x_L = 0.1257 f_{GHz} L_{nH}$
$B_L = \dfrac{0.159}{f_{GHz} L_{nH}}$	$[-j]$	$b_L = \dfrac{7.96}{f_{GHz} L_{nH}}$
Capacitors		
$X_C = \dfrac{159}{f_{GHz} C_{pF}}$	$[-j]$	$x_C = \dfrac{3.183}{f_{GHz} C_{pF}}$
$B_C = 0.006283 f_{GHz} C_{pF}$	$[+j]$	$b_C = 0.314 f_{GHz} C_{pF}$
Ideal lumped inductance and capacitance in nanohenries and picofarads		
$L_{nH} = \dfrac{0.159 X_L}{f_{GHz}} = \dfrac{0.159}{f_{GHz} B_L}$		$L_{nH} = \dfrac{7.96 x_L}{f_{GHz}} = \dfrac{7.96}{f_{GHz} b_L}$
$C_{pF} = \dfrac{0.159}{f_{GHz} X_C} = \dfrac{0.159 B_C}{f_{GHz}}$		$C_{pF} = \dfrac{3.183}{f_{GHz} x_C} = \dfrac{3.183 b_C}{f_{GHz}}$

Ideal stub reactance, susceptance, and electrical length

Open stubs

$X_{OS} = \dfrac{Z_{OS}}{\tan\theta}$	$\theta = \tan^{-1}\left(\dfrac{Z_{OS}}{X_{OS}}\right)$	$[+j]$	$x_{OS} = \dfrac{Z_{OS}}{Z_0 \tan\theta}$	$\theta = \tan^{-1}\left(\dfrac{Z_{OS}}{Z_0 x_{OS}}\right)$
$B_{OS} = \dfrac{\tan\theta}{Z_{OS}}$	$\theta = \tan^{-1}(Z_{OS} B_{OS})$	$[-j]$	$b_{OS} = \dfrac{Z_{0S} \tan\theta}{Z_{OS}}$	$\theta = \tan^{-1}\left(\dfrac{Z_{os} b_{os}}{Z_0}\right)$

Shorted stubs

$X_{ss} = Z_{ss} \tan\theta$	$\theta = \tan^{-1}\left(\dfrac{X_{ss}}{Z_{ss}}\right)$	$[+j]$	$x_{ss} = \dfrac{Z_{ss}}{Z_0} \tan\theta$	$\theta = \tan^{-1}\left(\dfrac{Z_0 x_{ss}}{Z_{ss}}\right)$
$B_{ss} = \dfrac{1}{Z_{ss} \tan\theta}$	$\theta = \tan^{-1}\left(\dfrac{1}{Z_{ss} B_{ss}}\right)$	$[-j]$	$b_{ss} = \dfrac{Z_0}{Z_{ss} \tan\theta}$	$\theta = \tan^{-1}\left(\dfrac{Z_0}{Z_{ss} b_{ss}}\right)$

Input impedance of an ideal transmission line of electrical length θ, terminated with Z_L

$$Z_{IN} = Z_{TL} \frac{Z_L + jZ_{TL} \tan\theta}{Z_{TL} + jZ_L \tan\theta}$$

Impedance and length of a cascade line to match $Z_S = (R_S + jX_S)$ to $Z_L = (R_L + jX_L)$

$$Z_{TL} = \sqrt{\frac{(R_S^2 + X_S^2)R_L - (R_L^2 + X_L^2)R_S}{R_S - R_L}}$$

$$\theta = \tan^{-1}\left[\frac{Z_{TL}(R_L - R_S)}{X_S R_L - X_L R_S}\right]$$

If $X_S = X_L = 0$

$Z_{TL} = \sqrt{R_S R_L}$ *and* $\theta = 90^\circ, 270^\circ, 450^\circ$

About the Authors

Rowan Gilmore is an experienced consulting engineer who introduced the world's first commercial harmonic-balance CAD simulator while he was vice-president of engineering at Compact Software. He has held numerous design and management posts in industry, including Central Microwave, Schlumberger, Telstra, and SITA. A senior member of the IEEE, he holds a D.Sc. and an MSEE from Washington University in St. Louis, and a B.E. in electrical engineering from the University of Queensland in Brisbane, Australia. He has nearly 15 years of teaching experience with Besser Associates and CEI Europe.

Les Besser is the chairman of Besser Associates, a continuing education organization. He is a Life Fellow of the IEEE, in which he has held various offices and received awards and recognition for past accomplishments. He holds a Ph.D., an M.S., and a B.S. in electrical engineering. Dr. Besser authored the first commercially successful microwave circuit optimization routine, COMPACT, and founded Compact Software (now part of Ansoft), a pioneer group in RF/MW CAE. A master lecturer, he is currently heading an organization dedicated to continuing education through instructor-led and Internet-based short courses, and CD and videotaped presentations. His company has trained nearly 50,000 engineers and managers since the mid-1980s.

Both authors may be reached at http://www.bessercourse.com.

Index

$\pi/4$ DQPSK modulation, 522, 524
μ-factor, 32–35
$1/f$ noise, 102
Active circuits, 1
  dc bias, 63–64
  design, 74
Active dc bias circuits, 63–64
  depletion-mode FETs, 63
  enhancement-mode FETs, 63
  in low-voltage circuit applications, 64
Active FET mixers, 479–83
  AFT-54143, 480–83
  analysis, 479
  example, 480–83
  harmonic-balance simulation, 482
  schematic, 479
  See also Active transistor mixers
Active multipliers, 502
Active RF devices, 147–89
  diode model, 148–50
  two-port device models, 150–89
Active transistor mixers, 464–88
  bipolar, 467–78
  defined, 464
  design principle, 465
  FET, 479–83
  Gilbert cell, 483–88
  input filters, 466
  See also Mixers; Transistor mixers
Active two-port
  device models, 150–89
  S-parameter, 31
  stabilizing, 46–50
  See also Two-ports
AD8347 receiver front end
  block diagram, 537
  defined, 536
  See also Integrated system chips
Adjacent channel power (ACP), 314, 513
Adjacent channel power ratio (ACPR), 305
  defined, 308
  deriving, 309
Admittance
  emitter load, 417, 418
  formula, 547
  matrix, 203, 204
  normalized, 416
AFT-54143 HEMT, 480–83
  circuit using, 481
  conversion gain, 483
  defined, 480
  drain current, 483
  IF output power, 483
  See also Active FET mixers
Air interface specification, 522–23
  modulation scheme, 522, 524
  receiver sensitivity, 524
  See also Personal handyphone system (PHS)
Amplifier design
  for maximum gain, 82–88
  quasi-linear, 223–43
  with single matching networks, 13–14
  single-sided, 15–19
  S-parameters, 2
Amplifiers
  balanced, 114–16
  bilateral, 78–88
  broadband, 123–42
  categories of, 243–80
  class-#, 275–78
  class-A, 227, 243–48
  class-AB, 267–68
  class-B, 248–57
  class-C, 268–69
  class-D, 271–75
  class-F, 257–65
  design for maximum gain, 82–88
  distributed, 141–42
  feedback, 129–41
  harmonically controlled, 269–71
  LINC, 326, 327
  linear, 77–142
  load line, 224–32
  maximum gain, 80

Amplifiers (continued)
multistage, 88–93
push-pull, 252–55
switching-mode, 271–78
Analog-to-digital converter (ADC), 515, 521, 522
Analytical methods, 194
AppCAD program, 68–69
downloading, 68
solution of dc bias circuit, 69
worst-case collector currents, 70
Arbitrary frequency multiplication, 505–6
AT-64020 bipolar transistor, 245–47
characterized data, 247
data sheet, 246
defined, 245
See also Class-A amplifiers
Automatic gain control (AGC), 311
Available gain design, 107–21
circles, 109, 111
defined, 107
LNA, 110–11
outline, 108–10
source termination selection, 108
Available power gain, 3, 6
AWT6200 PowerPlexer, 544
defined, 544
functional block diagram, 545
See also Integrated system chips
Balanced amplifiers, 114–16
advantages, 114–15
disadvantages, 115
layout, 120
noise figures, 116
summary, 116
Barkhausen criterion, 340, 357
Baseband digital processing, 518–20
Base bias resistance, 323
BAT17 Schottky diode pairs, 447–48
data sheet, 447
forward voltage, 448
LO power requirements, 448
BFP640 mixer, 470–78
collector current, 476
conversion gain, 477
defined, 470
illustrated, 475
operation, 475–76
output power, 478
output spectrum, 478
simulated large-signal input reflection coefficient, 474
simulations, 477
small-signal equivalent input circuit model, 473
small-signal input match, 471
small-signals, 470
See also Bipolar transistor mixers
BFR360
defined, 410
Gummel-Poon model for, 410
large-signal *S*-parameters, 412
simulation in common-collector configuration, 413
small-signal *S*-parameters, 411
Bias
base, resistance, 323
changes at input, 298–302
changes at output, 302–4
chokes, 69
considerations with power devices, 304–7
to control output power, 306–7
network to support high currents, 305–6
points, input/output, 303
stabilization, 306
temperature effects and, 306
voltage, 228, 230, 249
BiCMOS technology, 188
Bilateral amplifiers
block diagram, 79
design computations, 80
design for maximum gain, 82–88
design for maximum small-signal gain, 78–88
Bilateral design, 7, 19
Bipolar CMOS (BiCMOS), 543
Bipolar transistor mixers, 467–78
BFP640, 470–78
conversion gain, 469
example, 470–78
gain expression, 469
limitations, 469–70
principle, 467, 468
small-signal equivalent circuit, 468
See also Active transistor mixers
Bipolar transistors
applied voltages, 155
AT-64020, 245, 246, 247
breakdown effects, 161–63
breakdown voltage, 162
common-base configuration, 130, 172–73
common-collector configuration, 130
common-emitter configuration, 130
dc model, 156
Ebers-Moll circuit topology, 158
Ebers-Moll model, 153–61
gain, 60
Gummel-Poon model, 163–69
heterojunction, 173–77

input-output signal voltage, 130
knee voltage, 160
linear region, 160
model, 153–73
noise figure, 60
off region, 160
resistive negative feedback bias circuits for, 61
reverse collector current, 161
saturation region, 160
small-signal model (Ebers-Moll derivation), 163
small-signal model (Gummel Poon derivation), 169–72
stabilization of, 50–59
T-topology model, 157
Bode plot, 358
checking Nyquist stability criterion with, 358
comparison, 359
open-loop gain, 363
phase, 363
Pierce oscillator, 360
Boundary conditions, 223
Breakdown
effects, 161–63
limits, 162–63
onset of, 162
voltage, 162
See also Bipolar transistors
Bridged-T, 129
Broadband amplifiers, 123–42
amplifier-equalizer combinations, 129
bandwidths, 123
cost/performance, 128
dissipative mismatch at input/output ports, 125–29
distributed, 141
feedback, 129–41
noise figure, 138
reactive match/mismatch approach, 124–25
RF circuit schematics, 129
single-stage 800-to 2000-MHz, 126–29
Broadband stability analysis, 57–59
Butler oscillator, 385
CAD
commercial suites, 402
optimizer, 348
Capacitance
Colpitts, 380
grid-to-plate, 6
Miller, 380, 383
parasitic, 408
varactor, 390, 420–21
Capacitors
MOS varactor, 375
output blocking, 224
varactor, 389
Carrier-to-intermodulation (C/I) ratio, 312
Cartesian Loop Architecture, 326
Cartesian loop transmitter, 326
Cascade noise figure, 530, 548
Cascading amplifiers, 88–93
block diagram, 279
design, 278–80
by direct impedance matching, 89–92
gain, 92
impedance match, 92–93
output power, 92–93
power, 278–80
return loss, 92
cdma2000, 514
Ceramic resonator, 385
Chalmers model, 183
CHP1207-QM chip, 543–44
Circuit layout, 71–73
EM simulators for, 71
for parallel stabilizing branch, 72
Circuit optimization, 127
Circulator implementation, 211
Circulators, 521
Clapp oscillator, 379–80
definitions, 379
illustrated, 379
implementation, 379
net capacitance, 389–90
reactance of resonant load, 391
resonant circuit, 388, 391
VCO, 387
Class-A amplifiers, 227, 243–48
AT-64020, 245, 246, 247
comparison, 265–67
dc power, 244, 245
defined, 243
device voltage, 243–44
dissipated power, 245
efficiency, 244, 245
example, 245–48
impedance level, 255
load line comparison, 266
maximum efficiency, 244
output power, 244, 245
total device current, 243
zero-to-peak sinusoidal amplitude, 244
See also Amplifiers
Class-AB amplifiers, 267–68
defined, 267
efficiency, 268
I-V curves, 268

Class-AB amplifiers (continued)
operation, 267
Class-B amplifiers, 248–57
advantages, 250
biasing, 248
characterization of, 255–57
comparison, 265–67
conduction angle, 248
dc power, 250
defined, 248
gain, 250
harmonic components, removing, 252–53
input power, 266
input voltage, 248
load line, 252
load line comparison, 266
optimum load resistor, 251
output current, 248
output power, 250
output voltage, 249
push-pull configuration, 252–55
See also Amplifiers
Class-C amplifiers, 268–69
defined, 268
FETs, 505
mixed-mode terminations, 269, 270
precautions, 269
uses, 268
See also Amplifiers
Class-D amplifiers, 271–75
current-mode, 273
defined, 271–72
principle, 272, 273
voltage/current waveforms, 274
voltage mode, 272–73
See also Amplifiers; Switching-mode amplifiers
Class-E amplifiers, 275–78
defined, 275
deployment, 278
extrinsic drain load, 275
modeling, 275
output power, 277
output voltage, 275
topology, 276
waveforms, 275, 276
See also Amplifiers; Switching-mode amplifiers
Class-F amplifiers, 257–65
comparison, 265–67
defined, 257–59
drain current, 265
drain voltage, 265
efficiency, 259
example, 263–65
FET, 263
harmonic terminations, 273
input bias, 259
input power, 266
inverse, 263
layout, 264
load line comparison, 266
load line resistance, 264
odd-order harmonic components, 266
output resistance, 263
output voltage, 259
principles, 260
reduced device dissipation, 261
simulated waveforms, 265
theory limitations, 262
See also Amplifiers
Closed-loop system
Colpitts oscillator, 409
equivalent power spectral density, 396
illustrated, 339
oscillator analysis, 338–41
for oscillator modeling, 362
output, 396
resonator model, 395
signal modulation in, 395–400
Code division multiple access (CDMA), 511
ACP, 513
channel spacing, 518
defined, 513
multicarrier, 514
spreading code, 514
wideband (WCDMA), 512, 513–15
Colpitts oscillator, 376–78
analysis, 376, 377
broadband negative resistance, 388
capacitance, 380
closed-loop system, 409
configuration illustration, 376
crystal, 383
examples, 385–90
illustrated, 379
impedance, 384, 409
implementation, 379
input impedance, 377, 378
RF configurations, 384
terminated by R-L circuit, 384
variants, 378–80
See also Transistor oscillators
Colpitts oscillator design, 404–10
capacitive reactance, 407
crystal parasitic capacitance effect, 407–10
equivalent input circuit, 406
expected effective impedance, 408

illustration, 406
input *S*-parameters, 404
parasitic capacitance, 408
topology, 406
See also Colpitts oscillator design; Oscillators
Common-base configuration, 130, 172–73
Common-emitter transistors, 374
Common-source transistors, 374
Conductance
formula, 547
incremental, 435
negative resistance oscillator, 356
Constant-gain circles, 125
Constant-output-power contours, 97
Conversion gain, 438, 439, 469, 477, 483
Crossing angle, 365, 369
Crystal oscillators, 380–85
Butler, 385
Colpitts, 383
equivalent circuit, 381
input reflection coefficient, 381
motional arms, 380
overtone, 382
See also Transistor oscillators
Current-limited circuits, 228
Current-mode class-D amplifiers, 273
Curtice model, 180
cubic, 182
quadratic form, 181
dc bias, 118
active circuits, 63–64
AppCAD solution, 69
arrangements for FETs, 62
bipolar transistor gain/noise figure, 60
circuit design, 69
feeding, into RF circuit, 64–69
FETs, 118
network filtering, 69
parameters, finding, 66
passive networks, 60–63
resistive feedback circuit for, 68
resistive networks, 135
techniques, 59–69
worst-case analysis, 69–71
dc transfer characteristics, 66, 67
of BFP 640 transistor, 67
establishing, 66
Digital cordless systems, 511
Digital downconversion (DDC), 518
Digital signal processors (DSPs), 307
as linearizers, 328
processing speed, 522
Digital-to-analog converter (DAC), 515
Diode mixers, 442–64
double-balanced, 451–55
harmonic components, 460–64
image problem, 455–60
nonlinearity, 488
single-balanced, 445–51
single-ended, 443–45
topologies, 442–43
See also Mixers
Diode model, 148–50
low-frequency equation, 148
p-n junction, 148
stored-charge, 148
Diode predistorter circuits, 315
Direct impedance matching, 89–92
example, 90–92
highpass/lowpass interstage network, 91
interstage networks, 90
two-stage amplifier with, 90
Directional couplers
illustrated, 214
oscillator analysis using, 214–15
OSCTEST, 214
Direct upconversion, 519
Dissipative mismatch, 125–29
Distortion
amplifier linearity and, 309–11
components, 322
defined, 219
device modification and, 319–25
feedback cancellation and, 317–19
IMD, 307
linear devices and, 320
predistortion, 312–17
reduction, 307–28
relative output, 324
system-level reduction of, 325–28
third-order, 317
Distributed amplifiers, 141–42
defined, 141
input/output capacitances, 142
Doherty amplifier, 324
Double-balanced mixers, 451–55
circuit topology, 451
constructed from four dual-gate FET mixers, 499
defined, 451
diode embedding impedances, 452–53
EMD40-2400L, 453–54
illustrated, 455
phase relationships, 452
ready-made, 453
side arms, 451
third-order intercept point, 493–94

Double-balanced mixers (continued)
See also Diode mixers; Mixers
Downconverter system, 436, 437
Drain current, 236, 265, 304
average, 303
total, 304
Drain efficiency, 222
Drain terminals, 285
Drain voltage, 236
class-B amplifiers, 251
class-F amplifiers, 265
peak-to-peak, 237
Dual-gate FET mixers, 494–500
comparison, 501
defined, 494
double-balanced mixer constructed from, 499
I-V curves, 496, 497
LO power requirement, 498
operation, 494–95
principle, 495, 496
schematic, 496
in single-balance configuration, 499
See also Mixers
Ebers-Moll model, 153–61
applied voltage, 154, 155
base-emitter junction, 154
breakdown effects and, 161
collector-based junction, 154
dc model, 155, 156
drawbacks, 161
illustrated, 154
linear region, 160
off region, 160
recombination of electrons, 156
reverse injection, 155
saturation region, 160
small-signal transistor model derived from, 163
topology for bipolar transistor, 158
T-topology model, 157–58
Electromagnetic (EM) simulation, 43
2.5-D, 72
3-D, 72
circuit combining with, 72
circuit layout and, 71
EMD40-2400L, 453–54
data sheet, 453
defined, 453
schematic, 454
See also Double-balanced mixers
Envelope restoration techniques, 327
Equivalent noise resistance, 106
Equivalent series resistance, 368
Error vector magnitude (EVM), 519
Feedback
dual, 134
external passive circuits, 131
filters, 397
lossless, 50–51, 110, 324
negative, 129–30, 132, 138
positive, 130
series inductive, 51
transformers for, 324
unwanted, 25
Feedback amplifiers, 129–41
in broadband RF systems, 141
component tolerance effects, 140–41
design example, 134–40
design formulas, 133–34
design procedure, 133
open-loop gain, 131, 132
RF equivalent circuit of, 136
RF schematics, 139
RF simulations, 140
See also Amplifiers; Broadband amplifiers
Feedback loop, 138
adding loss into, 138
oscillator formed by closing, 346
Feedback resistors, 136
parallel, 137
series, 137
Feedforward cancellation, 317–19
advantages, 317
delay lines, 318
limitations, 319
principle, 317–18
suppression of, 319
FET mixers
active, 479–83
comparison, 500–501
dual-gate, 494–500
quad, 526
resistive, 488–94
using, 501
See also Mixers
FETs
amplifier, 329
class-C, 505
class-F, 263
CMOS, 188, 375
common-gate, 332
common-source, 332
dc bias arrangements for, 62
depletion-mode, 63
dual-gate, 316
enhancement mode, 62, 63, 116
with feedback, schematic diagram, 348

gate bias network, 299
in Gilbert cell topologies, 488
packaging, 493
stability circles, 296
as subharmonic mixers, 461
FHX35LG HEMT, 502–3
Finite impedance, 53
Flicker noise, 102
Frequency division duplex (FDD), 511
Frequency division multiple access (FDMA), 510, 511, 522
Frequency-domain techniques, 200–201
Frequency doublers, 502–5
balanced, with 90° couplers, 504
bandwidth, 504
conversion gain, 504
nonlinear behavior, 502
Frequency multipliers, 501–6
active, 502
arbitrary, 505–6
doublers, 502–5
overview, 501–6
passive, 502
Frequency pushing, 402
Frequency-shift keyed (FSK) modulation, 309
GaAs MESFETs, 177–84
Gain, 87
active device, 131
amplifier, 131
available, design, 107–21
class-B amplifiers, 250
conversion, 438, 439, 469, 477, 483
current, 170
equalization, 92–93
feedback circuits and, 137
feedback loop, 63
flatness, 138
forward, 340
LNA, 528
maximum, amplifier design for, 82–88
open-loop, 131, 132, 214–15, 342, 349
operating, 94–101
selective, compensation, 127
series inductive feedback effect on, 51
simulated, 120
small-signal, 99
transducer power, 2, 5, 6, 78
unwanted, dissipating, 126
Gain-bandwidth product, 170
Gate-bias voltage, 299, 300
Gilbert cell mixer, 483–88
defined, 483
RF drive, 488
schematic, 484
use, 483
See also Active transistor mixers
Gilbert cell multiplier, 488
Global positioning satellite (GPS), 516
GSM systems, 307, 512
Gummel-Poon model, 163–69
ac model parameters, 164–65
AF parameter, 402
base-width modulation, 164
complexity, 167
equivalent circuit topology, 167
forward current gain variation, 166
high level injection, 164
KF parameter, 402
low current effects, 164
modification to, 168
output resistance, 166–67
PNP, 168
problems, 168–69
small-signal model derived from, 169–72
See also Bipolar transistors
Gunn diode oscillators, 373
Harmonically controlled amplifiers (HCAs), 269–71
dynamic load line, 269
half-sinusoidal (hHCAs), 269
input voltage control, 270
rectangular (rHCAs), 269
Harmonically-tuned MESFET, 296–98
efficiency, 298
fundamental output power, 298
intrinsic drain voltage/current, 297
power-added efficiency, 298
simulated gain, 298
simulated response, 297–98
Harmonic balance, 197–200
advantages, 198–99
analysis of oscillators, 207–15
convergence, 207
error (HBE), 206
method, 202–7
principles, 204
priori assumption, 198
process steps, 204
simulators, 197
speed advantage, 200
time-domain techniques vs., 198
Harmonic terminations, 235
Harmonic tuning, 296–98
Hartley oscillator, 339, 379
illustrated, 339
with wideband tuning, 386
See also Oscillators

Heterojunction bipolar transistors (HBTs), 173–77
AlGaAs, 177
defined, 173
GaAs, 173, 174
Gummel plot, 174
InGaP, 173, 176
modeling, 173–77
SiGe, 175–77
T-model, 158
High-electron mobility transistors (HEMTs), 184–87
defined, 184
enhancement mode, 301
GaN, 186, 187
InP, 186
measured characteristics, 185
normal depletion mode, 185
performance, 184
pseudomorphic, 321
SiC, 186
High-power RF transistor amplifiers, 217–328
bias considerations, 298–307
categories, 243–80
design example, 280–98
distortion reduction, 307–28
nonlinear concepts, 217–23
quasi-linear design, 223–43
Ideal circuit elements, 58
Image problem, 455–60
defined, 455
solutions, 455–56
See also Mixers
Image-reject mixers, 440–41, 456–60
as form of complex mixing, 458
frequency-shifting properties, 459
path of unwanted image signal, 457
principle, 456
in reverse direction, 459
schematic, 457
structure, 459–60
See also Mixers
Impedance matching
cascading amplifiers, 89–92
direct, 89–92
problems, 92–93
Incremental conductance, 435
Inductive resonator, 379
Input matching circuit, 85
Input third-order intercept point (IIP3), 325
Instability
defined, 20
low-frequency, 70
multiband, 51
potential, 32, 41–42
from unwanted feedback, 25
See also Stability
Integrated system chips, 531–44
AD8347, 536, 537
AWT6200, 544, 545
CHP1207-QM, 543–44
MAX2291, 544
MAX2338, 533, 535
MAX2361/3/5 series, 538, 539
MAX2420, 538–40
MC13190, 540–41
MC13751, 536–38
power amplifier modules, 543–44
RF receiver front ends, 532–36
RF upconverters and transmitter driver amplifiers, 536–38
SA1920, 534
transceiver and complete radio solutions, 538–43
UAA3535HL, 541–43
International Telecommunication Union (ITU), 512
Interstage matching, 91
Interstage second harmonic enhancement, 317
Interstage stability analysis, 44–46
Intersymbol interference (ISI), 514
Isolation, 78, 255
I-V curves, 150–53, 165
defined, 150
dual-gate FET mixers, 496, 497
flat region, 150
illustrated, 151
knee, 150, 307
of MESFET, 181
output, 166
slope, 153
spacing, 151, 152
Johnson noise, 102
K-factor, 31–32
defined, 32–33
physical interpretations, 33
unstable terminations and, 33
Kirchoff's laws, 195, 206, 350
Lange coupler, 114
Laterally diffused MOSFET (LDMOS), 188
L-C oscillators, 337, 364, 374–76
Limiting, 78
LINC amplifier, 326, 327
Linear amplifiers, 77–142
categories, 77–78
design considerations, 1–74
low-noise, 77
maximum absolute output power, 77–78
maximum small-signal gain, 77
Linearity, 309–11

ATF-54143, 320
Doherty amplifier and, 324
of mixers, 441
Linearizers
DSPs as, 328
predistorter, 314
Load
bilateral effects, 236
impedances, 243
optimum, 241
optimum, locus, 242
point, optimum, 242
power, maximum, 237
range, 241
tuning, 369
Load lines, 224–32
calculating, 224
class-B operation, 252
as dc relationship, 225
defined, 224
dynamic, 265
for maximum power/gains, 231–32
optimum, 283
output, 226
output power calculation from, 227
output voltage, 226
simulated characteristics, 288, 293
slope, 225, 227
static, 265
trajectory constraints, 225
for transistor amplifier, 229
transmitter power amplifier, 531
of VCO, 428
Load pull
active, measurement system, 233
methods, 232–43
nonlinear simulations, 286
passive, measurement system, 233
tuners, 235, 297
Load pull contours, 233–34
approximate, 242
calculation transistor model, 239
collapse, 235
construction process, 238
creating, 238–43
defined, 233
illustrated, 240, 242
loci of, 234
predicting, 235–38
transformation of, 240
uses, 243
Load stability circles, 36, 51
Local oscillator (LO) ports, 433
Loop gain, 34, 348
Lossless feedback, 50–51, 110
effect, 50–51
series inductive, 51
See also Feedback
Lossy frequency selective gain shaping, 126
Low-frequency loop-gain filtering, 69, 70
Low-noise amplifiers (LNAs), 50
ADS schematics, 119
available gain design, 110–11
circuit schematic, 118
dynamic range, 106, 520, 522
gain, 528
multistage, 107
parallel, 117
schematic, 527
source selections for, 112
two-stage, 104
Lowpass matching networks, 133
Matching networks
added to output port only, 13–14
adding to input port only, 14
amplifier design with, 13–14
broadband, 294
finding, 15–16
input, element values, 85
lowpass, 133
topologies, choosing, 83
two-element, 84
two-element highpass, 19–20
two-section, 295
MAX2291 chip, 544
MAX2338 chip, 533, 535, 536
MAX2361/3/5 chips, 538, 539
MAX2420 chip, 538–40
defined, 538
functional diagram, 540
See also Integrated system chips
Maximum absolute output power, 77–78
defined, 77–78
operating gain design for, 94–101
See also Output power
Maximum linear output power, 122
Maximum oscillation frequency, 171
Maximum small-signal gain, 77, 122
bilateral amplifier design for, 78–88
defined, 77
Maximum stable gain (MSG)
computing, from *S*-parameters, 8
defined, 81
at frequencies, 81
MC13190 chip, 540–41
block diagram, 541

MC13190 chip (continued)
defined, 540
MC13751 dual-band upmixer, 536–38
Memory effects, 305
MESFETs
bias network, 282
Chalmers model, 183
cross-section, 177
Curtice model, 180, 181, 182
dc drain current for, 302
drain current/voltage, 225
equivalent circuit model, 179–80
extrinsic drain current/voltage, 291
GaAs, 177–84
with harmonic tuning, 296–98
high speed, 178
as horizontal device, 177
input resistances, 179
intrinsic drain current/voltage, 291, 297
I-V curves, 231
large-signal models, 180–83
linear model, 284
measured/simulated I-V curves, 181
NE6500379A, 255, 256
Ooi model, 183
Parker-Skellern model, 183
power, 178
simulated load line characteristics, 291
simulated load pull contours, 286
simulated swept power characteristics, 290
small-signal equivalent circuit, 183
small-signal model, 183–84
topology, 179
tuning with input/output impedances, 287
Microstrip line discontinuities, 119
Miller capacitance, 380, 383
Minimum noise, 122
Mismatch loss, 16, 17, 548
Mixed-mode terminations, 269, 270
Mixers, 433–501
baluns and, 434
broadband, 440
combiners and, 434
conversion gain, 438, 439
conversion loss, 438, 440
diode, 442–64
double-balanced, 451–55
downconverter system, 436, 437
equivalent of harmonic components, 441
followed by IF amplifier, 441
harmonic components, 460
illustrated model, 433
image problem, 455–60
image-reject, 440–41, 456–60
incremental conductance, 435
linearity, 441
LO ports, 433
modeling, as switches, 438
output current, 436
output intercept point of, 464
overview, 433–42
SFDR, 442
single-balanced, 445–51
single-ended, 443–45
spurious components, 461–64
subharmonic, 460–61
transistor, 464–501
upconverter system, 437
Mobile telephony systems, 509–15
first generation, 509–10
second generation, 510–11
third generation, 512–15
Models
bipolar transistor, 153–73
Chalmers, 183
Curtice, 180, 181, 182
dc, 156
diode, 148–50
Ebers-Moll, 153–61
Gummel-Poon, 163–69
MESFET equivalent circuit, 179–80
MESFET large-signal, 180–83
MESFET small-signal, 183–84
Ooi, 183
Parker-Skellern, 183
physics-based, 147
small-signal transistor, 163, 169–72
T-topology, 157–58
two-port device, 150–89
VBIC, 169
Motional arms, 380
Motor-boating, 63
Multistage amplifiers, 88–93
cascading impedance-matched stages, 88–89
direct impedance matching, 89–92
LNA, 107
for narrowband applications, 92
output power and impedance matching, 92–93
NE6500379A, 255–56
biased/tuned by load-pull tuners, 258
data sheet, 256
data sheet schematic layout, 289
defined, 255
measured gain, 282
power-added efficiency, 282
schematic, 284

Negative feedback, 129–30, 132
applying, 138
benefits, 140
See also Feedback; Feedback amplifiers
Negative impedance, 364–69
Negative resistance, 347
defined, 28
occurrence, 28
at output port, 347
series circuit, 211
shunt circuit, 211
Negative resistance oscillator, 210, 353–57
load conductance, 356
load resistance, 356
shunt type, 355
summary, 356
total load, 354
See also Oscillators
Neutralization, 7–8
applying, 7
defined, 7
full, 8
grid-to-plate capacitance, 6
partial, 8
unilateral design vs., 8
Noise, 102–7
equivalent, resistance, 106
flicker, 102
measure, 103
optimum, reflection coefficient, 106
performance, 138
phase, 387
in RF circuits, 102–7
shot, 102
sources, 102–6
temperature, 103
thermal, 102, 394
two-port, parameters, 106–7
Noise factor
defined, 102
formula, 548
overall, 104
Noise figure, 105
balanced amplifiers, 116
broadband amplifier, 138
cascade, 530, 548
cascade, calculations, 104–6
converting to, 104
defined, 103
formula, 548
simulated, 120
Nonlinear analysis, 426–29
Nonlinear circuit simulation techniques, 193–215
analytical methods, 194
classification of, 193–201
frequency-domain methods, 200–201
harmonic balance method, 197–200
time-domain methods, 194–97
Nonlinear devices
characteristic description, 218
concepts, 217–23
phenomena, 220–23
Nyquist filters, 523
Nyquist frequencies, 363
Nyquist plots, 45, 358
example, 363–64
of open-loop gain, 361
Nyquist rate, 205
Nyquist stability criterion, 358
checking, with Bode plot, 358
for oscillation startup, 212
Off-chip image filter, 529
One-port oscillator design, 349–73
crossing angle, 369
device behavior, 369
intersection point, 369
load impedance, 369
negative impedance characterization, 364–69
negative resistance, 353–57
Q factors, 369–73
series resonant circuit, 180–353
startup, 357–64
See also Oscillators
Ooi model, 183
Open-loop gain, 131, 132, 342, 361
Bode plot, 363
measuring, 214–15, 349
Nyquist plot, 361
parameters, 345
See also Gain
Open-loop oscillator design, 341–49
detail, 349
gain measurement, 349
input/output match, 343, 346
unilateral assumption, 344
Operating power gain, 3
bilateral circles, 95
circles, 100
defined, 6, 94
design approach, 94
design for maximum power output, 98–101
design output, 95–97
load termination, 94
for maximum linear output power, 94–101
source termination, 94

Operating power gain (continued)
stability considerations, 97–98
See also Gain
Optimum noise reflection coefficient, 106
Optimum source impedance, 106
Original equipment market (OEM), 98
Orthogonal frequency division multiplexing (OFDM), 310, 514
Oscillation
causes, 22–25
constant amplitude, 364
at customer site, 20
damage, 19
drawbacks, 22
embedded circuits for, 376
feedback-type, 24
fixing, 19
input loop, 33
low-frequency limit, 23
maximum, frequency, 171
output loop, 3
overall feedback loop, 34
startup, 215, 340
steady-state, 22
Oscillator analysis, 207–15
with directional coupler, 214–15
with probes, 208–9
with reflection coefficients, 209–14
Oscillators, 337–429
block diagram, 23
Butler, 385
Clapp, 379–80
Colpitts, 376–78, 385–90
configurations, 373–90
crystal, 380–85
design examples, 404–29
design principles, 338–404
Gunn diode, 373
Hartley, 339, 379, 386
L-C, 337, 364, 374–76
modeling with closed-loop feedback system, 362
negative impedance, characterization of, 364–69
negative resistance, 210, 353–57
one-port design approach, 349–73
output current, 337
phase noise, 390–404
Pierce, 359, 360, 378–79
Q factors, 369–73
R-C, 337
recast as feedback system, 361
reflection coefficients, 347
right-half-plane poles, 357
series resonant circuits as, 349–53
theoretical basis for design, 208
transistor, configurations, 373–90
tuning curve, 366, 423
tuning history, 366
two-port design approach, 338–49
VCO, 365, 410–29
Oscillator startup, 357–64
guaranteeing, 358
predicting, 212
Output power
class-B amplifiers, 250
class-E amplifiers, 277
dependence, 220
as function of varactor voltage, 428
fundamental, 235
harmonically-tuned MESFET, 298
input power vs., 232
less than maximum cases, 236
magnified plot, 223
maximum, 233
saturated, 232
total, dependence, 221
Output resistance, 227
Output third-order intercept point (OIP3), 325
Parallel resistive stabilization, 51–55
Parker-Skellern model, 183
Passive dc bias networks, 60–63
Passive multipliers, 502
PE4134 quad mixer, 493, 494
Peak power, 243
Personal handyphone system (PHS), 522
air interface specification, 522–23
defined, 522
See also PHS chip-set design
Phase
Bode plot, 363
modulation, 393, 395
slope, 363
Phase noise
baseline contribution to, 397
characterizing, 390–404
control of, 400–402
defined, 393
floor, 394
general expression, 393
impact on system performance, 403–4
minimum, design requirements, 400–401
at output, 396
performance, 387
plot, 399
read from spectrum analyzer, 394
reasons for controlling, 404
reciprocal mixing from, 403

signal modulation and expression for, 392–95
simulation of, 400–402
system specifications for, 403
as time jitter, 400
See also Oscillators
Phase-shift oscillators, 23
PHS chip-set design, 522–31
component design, 525–31
component specification, 523–25
IF upconverter design, 525–26
receiver considerations, 524–25
receiver design, 527–30
transceiver integrated circuit design, 526–31
transmitter considerations, 523–24
transmitter design, 530–31
See also Personal handyphone system (PHS)
Physics-based models, 147
Pierce oscillator, 359, 360
analytical plots, 360
circuit schematic, 359
from Colpitts configuration, 378–79
defined, 359
illustrated, 379
See also Oscillators
PLL loop filters, 543
Positive feedback, 130
Potential instability, 32
defined, 41
forms of, 42
graphical forms of, 41–42
Power-added efficiency, 222
harmonically-tuned MESFET, 298
NE6500379A, 282
Power amplifier modules, 543–44
AWT6200 PowerPlexer, 544, 545
CHP1207-QM, 543–44
MAX2291, 544
Power amplifiers
bias considerations, 298–307
design example, 280–98
harmonic tuning example, 296–98
input/output device matching, 286–96
quasi-linear, design, 223–43
stabilization, 280
transistor characterization, 282–86
transistor selection, 281–82
Power gain
available, 3, 6
definitions, 3–7
operating, 3, 6
transducer, 2, 5
Predistorters
diode circuits, 315
dual-gate FET, 316–17
interstage second harmonic enhancement, 317
linearizer, 314
self-phase distortion compensator, 315–16
Predistortion, 312–17
applications, 313
concept, 313
defined, 312
RF-type limiter, 313
See also Distortion
Probes
defined, 208
definition of, 208
introducing, 208
oscillator analysis using, 208–9
voltage/frequency, 208
Push-pull amplifiers, 252–55
with baluns, 255
principle of, 253
residual harmonic distortion, 255
symmetrical circuit, 254
use of, 254–55
See also Class-B amplifiers
Q factors
concept, 371
external, 371–72
loaded, 371
oscillator, 369–73
resonator, 372
Quadrature phase-shift keying (QPSK), 310
Quasi-linear power amplifier design, 223–43
amplifier load line, 224–32
load pull methods, 232–43
R-C oscillators, 337
Reactance
Clapp oscillator, 391
formula, 547
series resonant circuits, 352
VCO, 423
Reactive match/mismatch approach, 124–25
Receiver front ends, 532–36
Reflection coefficients, 87, 346
crystal oscillators, 381
equalized two-stage amplifier, 26, 27
formula, 547
load, 212
magnitude, 42
optimized base circuit for, 416
optimum noise, 106
of oscillator, 347
oscillator analysis using, 209–14
of single stage/equalized two-stage gain-module, 26

Reflection gain, 27
Residual sideband suppression (RSB), 460
Resistive FET mixers, 488–94, 488–94
  with 180° baluns, 493
  comparison, 501
  defined, 488–89
  deployment, 492
  equivalent circuit, 490
  PE4134, 493, 494
  principles, 489
  with quadrature coupler, 493
  schematic, 489
  third-order distortion products, 492
  See also Mixers
Resistors
  feedback, 136, 137
  minimum-loss, finding, 47–48
  optimum load, class-B amplifiers, 251
  parallel, 47
  series, 47
  stabilizing values, 55
Return loss
  formula, 547
  simulated, 120
Reverse leakage current, 306
RF choke, 225, 227
RF digital processing, 515–17
RF integrated circuits (RFICs), 142
RF/MW simulators, 71
RF receiver chain, 529
Root-locus plot, 358, 360
Root-mean-square (rms) terms, 393
SA1920 dual-band application circuit, 534
Schottky barrier diode, 149
Second generation mobile systems, 510–11
Second-order intercept points (IP2), 519
Selective gain compensation, 127
Self-phase distortion compensator, 315–16
  defined, 315
  FETs used for, 316
  See also Predistorters
Series feedback, 8
Series resistive stabilization, 55–57
Series resonant circuits, 349–53
  illustrated, 350
  load impedance, 370
  load reactance, 352
  output voltage, 350–51, 352
  time-domain response, 351
Shot noise, 102
SiGe HBT, 175–77
Single-balanced mixers, 445–51
  with 90° coupler as balun, 446
  AM noise cancellation, 445, 446
  defined, 445
  dual-gate FET, 499
  higher frequency implementation, 448
  hybrid coupler and, 446
  phase relationships, 445
  schematic, 445
  spurious components, 450
  transformer implementation, 445
  VSWR, 450–51
  See also Diode mixers; Mixers
Single-ended mixers, 443–45
  advantages/drawbacks, 444
  antiparallel diode pair, 444
  conversion loss, 463
  impedance match, 444
  LO drive level, 444
  principle, 443
  with short-circuited second/third-harmonic LO terminations, 462
  topology, 443
  See also Diode mixers; Mixers
Single-sideband (SSB) modulator, 459
Small-signal transistor model
  derived from Ebers-Moll model, 163
  derived from Gummel Poon model, 169–72
Software-defined radio, 515–22
  baseband (direct conversion) digital processing, 518–20
  goal, 515
  RF digital processing, 515–17
  transceiver issues, 520–22
  wideband IF digital processing, 517–18
Source stability circles, 36, 51, 53
  defined, 36
  of NE6500379A, 99
  See also Stability circles
S-parameters, 101
  amplifier design, 2
  bias dependency, 69
  bilateral procedures, 2–3
  common emitter, 82
  large-signal, 363
  large-signal common-emitter, 412
  large-signal transistor, 344
  measurement, 39
  of NEC NE6500379A, 98
  oscillator circuit, 361
  for quasi-line modeling, 410–12
  small-signal, 363
  small-signal common-emitter, 411
  stabilized, 82
  two-port, 5

SPICE model, 135, 214
Spurious-free dynamic range (SFDR), 442
Stability
  broadband, 34, 49–50
  dc, 68
  of different transistors, 33
  factor, 34, 52
  importance, 20
  K-factor, 24
  Nyquist criterion, 29, 212
  one-port, 25–30
  operating gain and, 97–98
  RF circuit, 19–46
  RF test, 31
  series inductive feedback effect on, 51
  two-port, 30–35
  unconditional, 31–32, 40–41
Stability analysis, 21
  with arbitrary source/load terminations, 25–30
  broadband, 57–59
  as first step, 49
  interstage, 44–46
  one-ports, 25–30
  of two cascaded transistors, 43
Stability circles, 35–40
  defined, 36
  FET, 296
  input/output, 295, 296
  interpretation of, 37
  load, 37, 51
  locations of, 48
  source, 36, 37, 51, 53, 99
  for stabilizing potentially unstable device, 48
  stable side determination, 38
  warning, 42
Stabilization
  active two-port, 46–50
  bias, 306
  bipolar transistor, 50–59
  dc, 61
  device, 51–59
  parallel, 47–48
  parallel resistive, 52–55
  of power amplifiers, 97
  resistive, 49
  series, 47, 48
  series resistive, 55–57
  small-signal gain and, 99
State variables, 194
Subharmonic mixers, 460–61
  FETs as, 461
  implementation, 460
  schematic, 460
Susceptance, 547
Sweet spot, 304
Switching-mode amplifiers, 271–78
  advantages, 271
  class-D, 271–75
  class-E, 275–78
  defined, 271
Temperature
  effects on bias design, 306
  noise, 103
Termination impedances, optimizing, 415–20
Terminations
  borderline, 35
  comparisons, 121
  friendly, 35, 38
  harmonic, 235, 273
  load, 84, 121
  mixed-mode, 269, 270
  source, 84, 121
  types of, 35
  unfriendly, 35, 38
  unstable source region, 38–40
Thermal noise, 394, 397
  additive effect of, 398
  voltage, 29
Thevenin equivalent network, 299
Thevenin's theorem, 300
Third generation mobile systems, 512–15
Third-order intercept point (IP3), 441, 526
Third-order intermodulation distortion (IMD3), 307
  calculation of, 521
  measurement of HBT, 322
  power measurement, 323
Time division duplex (TDD), 511, 520, 530
Time division multiple access (TDMA), 511, 522
Time domain
  analysis, 194
  methods, 194–97
  transmission line modeling in, 196
Time jitter, 400
Transconductance, 170
Transducer power gain, 2, 5, 78
  expression, 6
  finding, 5
  simultaneous conjugate, 80
  See also Power gain
Transistor mixers, 464–501
  active, 464–88
  comparison, 500–501
  dual-gate, 494–500
  resistive, 488–94
  See also Mixers

Transistor oscillators, 373–90
Clapp, 379–80
Colpitts, 376–78, 385–90
crystal, 380–85
design art, 374
Hartley, 379, 386
L-C topologies, 374–76
Pierce, 378–79
See also Oscillators
Transmitter power amplifiers
load line of, 531
output power spectrum, 532
T-topology model, 157
Tuning curve, 366, 423, 428
Two-port device models, 150–89
bipolar transistor, 153–73
GaAs MESFET, 177–84
heterojunction bipolar transistor, 173–77
high-electron mobility transistor, 184–87
output terminals, 150–53
silicon LDMOS/CMOS technologies, 187–89
Two-port oscillator design, 338–49
closed-loop system analysis, 338–41
open-loop design, 341–49
See also Oscillators
Two-ports
applied input power, 94
bilateral, 6
cascading impedance matched, 88
frequency-dependent stability factor, 81
generalized block diagram, 3
input/output reflection coefficients, 4
noiseless, 107
noisy, 106
potentially unstable, 32, 35
RF stability test, 31
S-parameters, 1, 5
Two-port stability, 30–35
μ-factor, 32–35
K-factor, 31–32
potential, 32
unconditional, 31–32
Two-stage amplifiers
circuit schematics, 26
frequency response, 91
gain, 92
LNA, 104
reflection coefficient, 26, 27
return loss, 92
three matching networks of, 90
See also Amplifiers
Tx/Rx switch, 521
UAA3535HL transceiver, 541–43
application circuit for, 542
defined, 541
See also Integrated system chips
Unconditional stability
defined, 41
graphical forms of, 40–41
two-port, 31–32
See also Stability
Unilateral constant gain circles, 15–19
for finding matching network, 15–16
five constant-gain source circles, 17
illustrated, 16
Unilateral design, 2, 6–7
as approximation, 13
defined, 8
neutralization vs., 8
Unilateral figure of merit, 10–11
defined, 10
values, 11
Unilateral gain, 8–19
calculations, 12–13
constant circles, 15
defined, 8
maximum, 9, 12
See also Gain
Universal Mobile Telecommunications Service (UMTS), 512
Upconverter system, 436, 437
Varactor diode, 421
series resistance, 421
tuning bandwidth and, 422
Varactor voltage, 422, 425
frequency vs., 427
fundamental output power as function of, 428
load line of, 428
oscillator output spectrum with, 427
Vertical bipolar inter-company (VBIC) model, 169
Voltage-controlled oscillators (VCOs)
with Clapp configuration, 387
derived from nonlinear simulations, 428
device line tuning, 365
dynamic load line, 427
integrated, 369
noise, 399
off-chip, 543
tuning curve, 423, 428
tuning range, 437
varactor-tuned, 401
wideband, 374
Voltage-controlled oscillator (VCO) design, 410–29
analysis, 424–25
illustrated, 423
input reactance, 423

loading in unstable region, 412–15
nonlinear analysis, 426–29
quasi-linear modeling, 410–12
termination impedances, 415–20
tuning, 420–23
Voltage-limited circuits, 228
Voltage-mode class-D amplifiers, 272–73
Voltage standing wave ratio (VSWR), 547
Volterra kernels, 200, 201
Volterra series approach, 200–201
Wideband CDMA (WCDMA), 512
impact on RF design, 513–15
multicarrier, 514
multiuser signals, 517
systems, 513
Wideband IF, 517–18
Worst-case analysis, 69–71